Psychometrika

A JOURNAL DEVOTED TO THE DEVELOPMENT OF PSYCHOLOGY AS A QUANTITATIVE RATIONAL SCIENCE

THE PSYCHOMETRIC SOCIETY - ORGANIZED IN 1935

VOLUME 21
NUMBER 4
DECEMBER
1956

PSYCHOMETRIKA, the official journal of the Psychometric Society, is devoted to the development of psychology as a quantitative rational science. Issued four times a year, on March 15, June 15, September 15, and December 15.

DECEMBER, 1956, VOLUME 21, NUMBER 4

Published by the Psychometric Society at 1407 Sherwood Avenue, Richmond 5, Virginia. Second-class mail privileges authorized at Richmond, Virginia. Editorial Office, Department of Psychology, The University of Chicago, Chicago 37, Illinois.

Subscription Price: The regular subscription rate is $14.00 per volume. The subscriber receives each issue as it comes out, and, upon request, a second complete set for binding at the end of the year. All annual subscriptions start with the March issue and cover the calendar year. All back issues but two are available. Back issues are $14.00 per volume (one set only) or $3.50 per issue, with a 20 per cent discount to Psychometric Society Members. Members of the Psychometric Society pay annual dues of $7.00, of which $6.30 is in payment of a subscription to *Psychometrika.* Student members of the Psychometric Society pay annual dues of $4.00, of which $3.60 is in payment for the journal.

Application for membership and student membership in the Psychometric Society, together with a check for dues for the calendar year in which application is made, should be sent to

CHARLES WRIGLEY
Dept. of Psychology
University of California
Berkeley 4, Calif.

Payments: All bills and orders are payable in advance.

Checks covering membership dues should be made payable to the *Psychometric Society.*

Checks covering regular subscriptions to *Psychometrika* (for non-members of the Psychometric Society) and back issue orders should be made payable to the *Psychometric Corporation.* All checks, notices of change of address, and business communications should be sent to

DAVID R. SAUNDERS, Treasurer, Psychometric Society and Psychometric Corporation
Educational Testing Service
P.O. Box 592
Princeton, New Jersey

Articles on the following subjects are published in *Psychometrika:*

(1) the development of quantitative rationale for the solution of psychological problems;

(2) general theoretical articles on quantitative methodology in the social and biological sciences;

(3) new mathematical and statistical techniques for the evaluation of psychological data;

(4) aids in the application of statistical techniques, such as nomographs, tables, work-sheet layouts, forms, and apparatus;

(5) critiques or reviews of significant studies involving the use of quantitative techniques.

The emphasis is to be placed on articles of type (1), in so far as articles of this type are available.

(Continued on the back inside cover page)

Psychometrika

CONTENTS

VOLUME TWENTY-ONE DECEMBER 1956 NUMBER 4

Erratum in Rozeboom, W. W. and Jones, L. V., The validity of the successive intervals method of psychometric scaling. *Psychometrika,* 1956, **21**, 165–183.

On page 183, last sentence in first paragraph should read:

> "Essentially, what our present analysis has shown is that it is always possible to give a distributional definition to a base scale regardless of whether or not a scale exists which simultaneously normalizes all distributions."

PSYCHOMETRIKA—VOL. 21, NO. 4
DECEMBER, 1956

A TRIBUTE TO L. L. THURSTONE*

Louis Leon Thurstone's contributions to the development of psychology as a quantitative rational science are among the major scientific achievements of the present century. Many of us here still remember the thrill of learning about his early developments in psychophysical scaling methods during the 1920's. The psychophysical measurement methods developed in the latter part of the nineteenth century were used to measure the functional relation between the physical intensity of a stimulus, such as a light or a sound, and the psychological intensity of the sensation—the brightness of the light, the loudness of the sound. Thurstone was among the first to point out that these methods could be modified to construct a scale for measuring psychological qualities that had no measurable physical correlate. He developed the law of comparative judgment and demonstrated that in conjunction with the method of paired comparisons it could be used to measure purely subjective attributes, such as the aesthetic merit of paintings or the strength of an attitude.

In all his work, he stressed the fact that as long as we have merely a rule of procedure for analyzing data, we have no science. He insisted that every theory must be so precisely stated that one of the possible conclusions would be that the data collected were in disagreement with the theory.

In particular for the law of comparative judgment, he developed the criterion of internal consistency for psychophysical scales. For linear scales this means that the distance from object i to object j (as determined from one set of judgments "i greater than j") plus the distance from j to k (as determined from *another* set of judgments "j greater than k") should give a sum in reasonable agreement with the distance from i to k (as independently given by judgments "i greater than k"). Such a statement must be true for all possible sets of three, so that if one has as many as ten or twenty objects in the scaling experiment the number of checks becomes very large.

This, of course, is merely one illustration of the application of the criterion of internal consistency. One of Thurstone's important contributions was his insistence that each experimental and analytical procedure must contain such internal checks.

In addition to devising the theoretical framework for such psychological measurement, we remember Professor Thurstone for numerous applications of these methods to practical problems that were carried out by him and

*A statement read by Harold Gulliksen on behalf of the Psychometric Society at the annual meeting of the Society, September 4, 1956.

his students. One of the early applications was to measurement of the effects of various movies on attitudes of children; other studies measured the effect of various propaganda devices in changing attitudes, or measured the changes in national opinions as reflected in newspapers over a period of years. This area of psychological scaling opened by Thurstone over thirty years ago is still developing both in its theoretical aspects and its uses. Applications of these methods to measurement of intensity of attitudes and to the precise comparison of value systems have been or are being made. The psychophysical scaling methods are important and powerful scientific tools. During the coming decades, prolific and fruitful use of these methods in the development of such fields as linguistics, sociology, cultural anthropology, political science and economics will probably be seen.

Thurstone's achievements in psychology cannot be properly appreciated unless seen against the backdrop of psychological developments of the last half-century. Today it is taken for granted that aptitude and achievement tests can predict various types of academic and job performance with a useful degree of accuracy. Throughout this country, some hundreds of thousands of persons take some millions of tests annually. How often do we stop to remember that prior to 1900 there were literally no aptitude tests available for prediction of academic or job performance? Prediction of educability of children was one of the critical unsolved problems of democratic society. Commissions were appointed in various countries to study the problem. The Binet test was developed in France and was useful in predicting school achievement but was an unwieldy and cumbersome instrument. During the First World War, the Army Alpha and Army Beta tests were developed and used in the first large-scale mass testing program in the history of the world. L. L. Thurstone, as a young psychologist, was active in this testing program.

Thurstone saw the inadequacies in the then widely accepted notions of "general" intelligence. He realized that new and more powerful methods must be developed to analyze the masses of data necessary for a thorough study of the different aspects of cognitive ability, or as we now say, the different factors, or primary mental abilities. He made a very simple change in the then current theory. Instead of assuming that each person was to be characterized by *one* number G (his general intelligence), Thurstone assumed that it might take a great many numbers to describe the person—one number for each of the primary mental abilities.

He outlined his development of this problem to mathematicians whom he knew and thus learned that a field of mathematics called "matrix theory" existed. At over forty years of age, he decided to master a new field of mathematics, because it might help him in analyzing the nature of human abilities. He tutored regularly, worked problems, studied different texts—the result was the development of the factor methods that have been applied extensively by Thurstone, his students, and others, not only in analyzing the domain of mental ability but also in studying blocs in a legislature, schools of thought among teachers regarding curriculum content, classifying allergies, analyzing anthropometric measurements, organizing psychotic symptoms, and so on. The factor methods are extremely powerful, and opportunities for their application in the study of human behavior seem to be limitless. These developments and applications stem in large measure from Thurstone's gift for seeing an important problem, defining it clearly, and then sparing no effort in his persistent search for a solution.

This development of testing over the past fifty years should be seen not only as a scientific achievement but also as a humanitarian accomplishment. Fifty years ago any young person who wished to enter on a given course of study had no alternative but to try, if he were permitted to, for months or even for years; eventually by trial and error he would succeed or be discouraged by repeated failures and cease trying. Now the inept student need not face the discouragement of tackling too high a goal, and the gifted student need no longer be dependent on attracting the attention of some influential person in order to obtain opportunity for advanced training. The unknown person of talent can be identified and encouraged to proceed with advanced training in some appropriate area. Such identification and utilization of various degrees of talent can be of enormous benefit to society and also a boon to the individual who derives added happiness from engaging in an occupation which matches his abilities.

Fifty years ago no aptitude tests were available for such prediction of various aspects of human performance. That we can now see our way reasonably clearly to this goal is in no small measure due to the theoretical and the practical contributions of Professor Thurstone to the field of psychological measurement. His outstanding contributions to the fields of factor analysis and psychophysical scaling by no means exhaust the range of his achievements toward developing psychology as a quantitative rational science. He worked on developing personality measures, and on a quantitative rational learning theory. His Ph.D. thesis of 1916 constitutes one of the early attempts to develop equations of the learning curve.

On his retirement from the University of Chicago, his work continued to be as productive and ingenious as ever. At the University of North Carolina he was leading projects concerned with developing new psychophysical scaling methods and developing a set of novel ideas for obtaining objective measurements of personality characteristics. It is to be hoped that these ideas will not be lost, but that some research worker or group of research workers will develop and validate personality tests along the lines indicated by Thurstone.

During more than a quarter century at the University of Chicago, in addition to making outstanding contributions in many scientific areas Professor Thurstone trained large numbers of students who came not only from the United States but also from various foreign countries to study with him—to learn the techniques he was developing, the general principles of scientific investigation, and the principles of the quantitative rational approach which he espoused.

Louis Leon Thurstone has a unique position among psychologists of this century as an original research worker and as an inspiring teacher. Many of us who were privileged to know him closely found him a helpful and understanding friend.

We, the members of the Psychometric Society, feel it our particular privilege to pay homage to his memory. He conceived of this Society and its journal, *Psychometrika.* To his efforts more than to those of any other individual, both the Society and *Psychometrika* owe their present status and even their very existence.

As psychologists, we feel that our past achievements and our future aspirations in the theory and practice of psychometrics have been greatly influenced by Thurstone's developments, by his insights, and by his standards. We feel ourselves pledged to further his ideal: the development of psychology as a quantitative rational science. The greatest honor we can accord him as an outstanding scientist lies in our resolve to continue the development of psychology in the rigorous tradition he did so much to establish.

PSYCHOMETRIKA—VOL. 21, NO. 4
DECEMBER, 1956

THE SCALE GRID: SOME INTERRELATIONS OF DATA MODELS*

CLYDE H. COOMBS†

UNIVERSITY OF MICHIGAN

Perhaps an appropriate subtitle for this paper would be "Some Speculations on the Interrelations of Psychological Methodologies." The *Scale Grid* is a name I have given to a model which presumes to define the underlying continuities between such diverse areas as psychophysics, objective testing, attitude studies including questionnaire and interview techniques, learning experiments, rating scale methods, essay examinations, and projective instruments. The intent of the Scale Grid is to make explicit the fundamental similarities and differences of the methodologies among these various areas of psychological research. The increasing abundance of models and methods in all of these areas, with their associated nomenclature and specialized vocabularies, makes a unification of them increasingly desirable.

In some of these areas of psychological research, serious and intensive efforts have been made to construct models on a genotypic level to explain and predict manifest behavior. One thinks here, for example, of the area of signal detection in psychophysics and of objective test performance. In other areas the models are less explicit and tend to be on a literary level. However, even in these latter areas one may look at the methods of analyzing data. Because there is always a model, at least implied, some of the elements of the models are evident. One may look at this universe of models, explicit and implicit, abstract certain universal elements, and try to characterize them.

Our starting point will be to determine the primitive datum in psychology. What we want is an abstract definition which will hold for every type of psychological observation. Let us begin by taking some examples, seeing what the basic abstract datum is in each case, and then formulating a general definition. In some psychophysical experiments, for example, individuals judge which one of several stimuli is the greatest. In view of what the experimenter subsequently does with the data, it is evident that he thinks of each stimulus as a point; the judgment of the individual is inter-

*Presidential Address to the Psychometric Society, September 4, 1956.

†I wish to express my deep appreciation of the courtesies, assistance, and critical audience furnished me in the preparation of this paper by Professor H. C. J. Duijker and Professor A. DeGroot of the University of Amsterdam, while I was a Fulbright Research Fellow there in 1955-56.

preted as an order relation on these points. Over many replications, the model may deal with a distribution of points for each stimulus and be a probability or actuarial model, but this is not the type of distinction which is at all important to us now. Whether the models are deterministic or actuarial is not relevant to the fundamental distinctions between methodologies I want to make. In fact, I want only to make the point now that in some types of experiments the manifest behavior is interpreted as an order relation between a pair of points, both of which are identified with stimuli.

A different case arises, however, when an individual takes a mental test and passes some items and fails others. The use to which the behavior is put also suggests that it is being interpreted as a relation on a pair of points, but here one point is identified with an item and the other point with the individual. The point associated with the individual represents a measure of his ability; the point associated with the item represents its difficulty. The behavior of the individual in passing or failing the item is interpreted as an order relation on this pair of points. I am not here concerned with the numerical scores on a test or even how a theory arrives at such a score from the basic datum. These are differences on a higher level, and I am here concerned only with differences on most the primitive quantitative level.

When an individual is given an attitude scale and asked to indicate which items he will indorse, again the behavior is interpreted as a relation on a pair of points—one a stimulus, the other an individual. Here the relation is on a psychological distance between the point associated with an individual and the point associated with the stimulus. If the point associated with the stimulus is "near," in a sense defined by the model, the point associated with the individual, he indorses the item, otherwise not. So the behavior is interpreted as indicating whether the distance between two points is greater or less than a certain amount.

When an individual is asked to place a stimulus on rating scale, again the behavior is interpreted as a distance between the point associated with the stimulus and a point associated with a response category which is just another stimulus. Consider, for example, an individual who is asked to rate a stimulus, say a picture, as to whether it is superior, good, or poor. The picture is conceived of as being a point on the scale for this individual; the three points on the rating scale, superior, good, and poor, are also three stimulus points. From the latter points the individual selects the one nearest the point corresponding to the picture. So we see that rating scale behavior is interpreted as a relation between points. The same analysis holds if the rating scale is an ordered set of numbers or the real line. In fact counting is also rating scale behavior. If we ask an individual how many students there are in a class it does not matter whether he guesses or counts as far as the basic datum is concerned. The response, e.g., "35," is interpreted as a relation between one stimulus, the size of the class as perceived, and another stimulus, a real number.

When an individual is asked whether he observed a light increment or not, the behavior is interpreted as a relation between a point identified with the individual, a threshold, and a point identified with the stimulus, the magnitude of the increment. For a final example, consider an individual asked to judge which of two pairs of stimuli is more similar. Here the behavior is usually interpreted as a relation between two distances; if each distance is interpreted as a point, then behavior implies a relation on a pair of points.

All of these examples illustrate one important fact: behavior is made into data by interpreting it as an order relation between points or a relation on distances—both may more generally be regarded as a relation on a pair of points. An important distinction must be drawn between behavior and data. A datum is defined in this paper as a relation between points. That this is not a new idea is evident from a half-page note by Madison Bentley (1) in which he speaks of Stumpf, Wundt, Ebinghaus, Mach, G. E. Müller, and others who took the view that psychological measurement is a distance measurement, which is just a special case of a relation between pairs.

We have been speaking here as if there were just a single distinction between behavior and data. Actually a threefold distinction should properly be made. We may use the term *behavior* to refer to anything observable about the individual, *raw data* to refer to that which is selected for analysis, and *data* to refer to the interpretation of the *raw data* as an abstract relation between points. The first step in going from behavior to raw data, deciding what to observe, is a many-faceted problem which lies outside the scope of this paper. We are here concerned exclusively with the raw data and how it is interpreted as data in the sense defined above. I shall pursue this distinction between the *raw data* and the *data*, illustrating it in detail shortly.

In principle one could put any data in a matrix as follows: If the data were a relation on a pair of stimuli then a square matrix with rows and corresponding columns identified with stimuli (cf. Figure 1) could nicely accommodate the data. Each cell would contain an entry indicating the relation between that corresponding pair of stimuli. Another experiment, where one member of the pair of points was identified with an individual and the other with a stimulus would require that the matrix of Figure 1 be expanded as in Figure 2. Thus an experiment in which one point of a pair of points was identified as a stimulus and the other as an individual would be entered in the left portion; another experiment in which the behavior was interpreted as a relation on a pair of points, both of which were identified as stimuli, would be entered in the right portion.

Now if we go a little further and consider an experiment where the members of the pair of points are both identified with individuals, the matrix of Figure 2 becomes as in Figure 3. According to this figure, the behavior observed in some experiments is interpreted as a relation on pairs of points in which both points may be identified with stimuli, both points may be identified with individuals, or one point with a stimulus and one point with an individual. When the data are relations between stimulus points, the analysis yields the order of these points on a psychological attribute or the location of these points in a multidimensional space; the individuals who made the judgments or responses are not located as points in the space. So we might call such a space a *Stimulus* space. Correspondingly, we can talk about a space in which only individuals are located as a *Population* space, and a space with both stimuli and people as a *Joint* space. We have the kinds of spaces in which psychological data are analyzed classified in Figure 4.

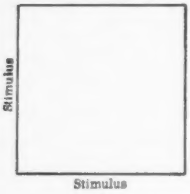

FIGURE 1

A Data Matrix

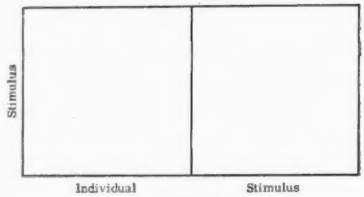

FIGURE 2

A Data Matrix

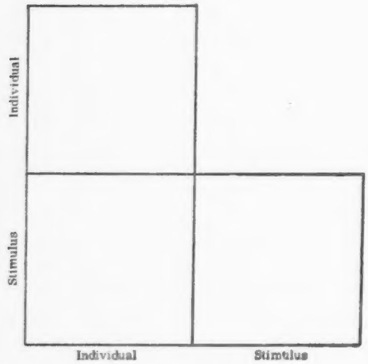

FIGURE 3

A Data Matrix

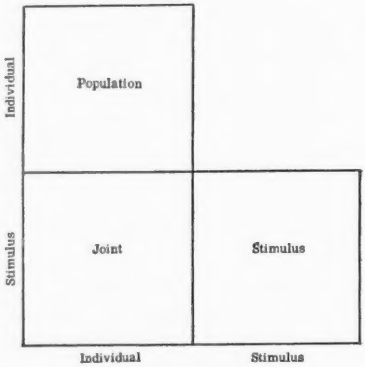

FIGURE 4

A Data Matrix

We now have the beginning of what I call the *Scale Grid*. We could move in either of two directions: developing the model and putting a little more meat on the skeleton, or constructing a psychological interpretation in order to bring out the implications of the grid. I find it rather difficult not to do both as they are so closely interdependent. So I shall first develop some of the theoretical ideas which will be most related to the interpretations which will follow. We shall consider some typical experiments which are mapped into Joint spaces and some which are mapped into Stimulus spaces; then we shall see what the difference is between them. The characterization of this difference will constitute one dimension of the Scale Grid.

Model Underlying The Scale Grid.

Behavior on a mental test is interpreted as a relation on a pair of points, one of which is associated with the individual and the other associated with the item or stimulus. Such behavior is mapped into data which, when analyzed, yields a Joint space with both stimuli and individuals located in it. What is the primitive operation here? The test item was conceived of as having a certain difficulty and the individual was, in effect, asked to compare his ability level with the difficulty level of the item. In a psychophysical study of the thresholds of individuals, the same is true, e.g., the individual is asked whether he perceives the stimulus, *yes* or *no*, and the behavior is interpreted as a relation on a pair of points, one of which is associated with the threshold of the individual, the other with a stimulus magnitude.

On the other hand, in a psychological study designed to measure heaviness of weights, length of lines, brightness of lights, or what have you, what is the primitive operation? The stimuli are conceived of as points on an attribute continuum, and the individual is asked which of the stimuli is greater. The behavior is interpreted as data on pairs of points both of which are associated with stimuli. Analysis of the data locates the stimuli on a scale, but no attempt is made to locate the individual as a point on the scale. The result is a Stimulus space.

Suppose I have some attitude statements about the church. I want to scale the items and then measure people's attitudes with them. The first step is to scale the items, so we ask individuals to evaluate the items as to which is more pro-church. A data matrix is constructed which is analyzed to yield a Stimulus space, say a one-dimensional scale, with the items located as points on a continuum. We note that the individuals were asked to evaluate the items with respect to where the items were on the continuum and not with respect to any *point* on this continuum which corresponded to the individual. Having scaled the items, we turn around and ask the individual which items he indorses. When we analyze these data, we end up with the individuals located on the same continuum, because this time the experimenter gave the individuals a different task to perform. The individuals were each asked to evaluate the items with respect to some point on the continuum corresponding to his own attitude toward the church, so now we are in a Joint space.

I could go on with examples from conditioning experiments or studies in perception, etc., but will not take the time for it. The important thing is to see the essential difference between behavior which is mapped into a Joint space and behavior which is mapped into a Stimulus space. In all the experiments, there are always both individuals and stimuli—what is it that determines whether an experimenter maps his experiment into a Stimulus space or into a Joint space?

If you go back and look at experiments with this question in mind, it becomes obvious that the experimenter puts his experment in a Joint space or a Stimulus space according to whether he regards the individual as having evaluated the stimuli with respect to a point corresponding to himself, the individual, or whether the individual evaluated the stimuli with respect to an attribute. I have called these two kinds of tasks, task *A* and task *B*, respectively. Task *A* may be described as evaluative, having to do with the relation of stimuli to the individual himself. Task *B* may be described as substantive, having to do with the nature of the stimuli per se.

I formalized this distinction between task *A* and task *B* in the following manner. In all experiments, both the individuals and the stimuli are points in a space, but in task *A* the points associated with the individuals are independent of the points associated with the stimuli. Whereas in task *B*, where the individual is evaluating stimuli with respect to an attribute, the point associated with an individual is completely dependent upon the points associated with the stimuli he is evaluating. I will not try to go further with this now, but essentially what we have is one dimension of the Scale Grid with just its two extremes represented—complete independence of the individual's point from those of the stimuli and complete dependence of his point on those of the stimuli, corresponding to task *A* and task *B*, respectively.

I have taken a good deal of time just to give an intuitive notion of one dimension of the Scale Grid. Let me briefly say just a few words about a second dimension. We can ask exactly the same questions about the difference between a Joint space and a Population space; we would find an equivalent answer if we just reverse the roles of stimuli and individuals in the argument and analysis just made. The reasoning is not difficult but is too detailed for an address. Let me merely state the conclusions. In a Joint space the points associated with the stimuli are completely independent of the points associated with the individual, whereas in a Population space the points associated with the stimuli are completely dependent on the points of the individuals responding to them. There is a duality between Stimulus spaces and Population spaces: in Stimulus spaces the points associated with individuals are dependent upon the points associated with the stimuli judged, whereas the reverse holds for Population spaces.

We have here two of the dimensions of the Scale Grid. I have constructed two others which can be used to characterize the data within these major areas of Joint spaces, Stimulus spaces, and Population spaces; but I will say nothing further about them here as they are not relevant to the interpretations I wish to bring out. I will only say that something of their nature is described in my early monograph on theory of psychological scaling (**2**).

If we take the two dimensions which we already have, they suggest a fourth type of space, called a *Field* space, in which the points associated with stimuli and those associated with individuals are completely mutually dependent. This gives us a two-dimensional Scale Grid as illustrated in Figure 5.

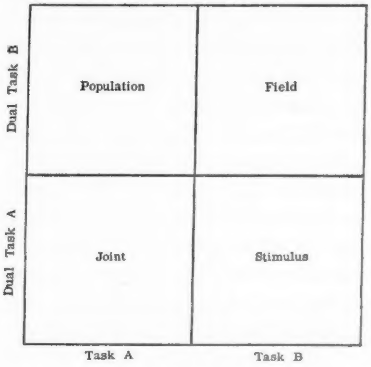

FIGURE 5

The Scale Grid

Some questions naturally arise as to just what might be the significance of all of this and just what this Field space type of behavior is. While we know what goes into a Stimulus, Population, or Joint space, this Field space was a consequence of our analysis of the others, and it is not immediately obvious what significance it has. In order to answer this question, I shall suggest a psychological interpretation of the Scale Grid.

Psychological Interpretation of the Scale Grid

I shall first point out what I consider to be the psychological processes involved in collecting data in a Stimulus space. Then by virtue of the duality between stimuli and individuals, a dual interpretation for Population spaces will be made and certain implications of duality pointed out. From these two kinds of spaces we shall move in one direction and get Joint spaces and in the other direction to get Field spaces.

Consider what is involved when we collect data in a Stimulus space. An individual, the subject, is asked to make judgments about stimuli with respect to an attribute (task *B*). He is given a set of weights and asked about their felt-heaviness, a set of tones and asked about their pitch or loudness, etc. The objects of judgment in this situation have many measures; each object has a measure on each of its many attributes—e.g., color, size, heaviness, form, volume, aesthetic quality, etc.

So we have stimuli corresponding to points having several components, and individuals instructed to select one of these components and evaluate the stimuli with respect to that attribute. Let me say it again and contrast the difference—the stimuli are points on many attributes, the individual comes with an attribute in mind but no scale position of his own from which to evaluate the stimuli. In an exaggerated sense, for a Stimulus space the stimuli have provided the points, and the individual has provided the attribute. These then are the respective functions of stimuli and individuals in generating data in a Stimulus space. The behavior observed is ultimately converted to measures of the stimuli on the attribute. The behavior observed may run the gamut of paired comparisons, rating scales, or free-answer protocols—these differences are not relevant here. The important thing is that the behavior observed is interpreted as data which are relations on the stimuli with respect to the attribute. The analysis then leads to measures of the stimuli only.

Let us now exercise the duality relation and consider what the process must be for a Population space. When we reverse the roles of stimuli and people we must have a group of individuals, each possessing measures on many attributes just as the stimuli had in a Stimulus space. An individual thus corresponds to a point with many components. Then we must have stimuli which, carrying through the analogy, must be instructed to select one of these components and evaluate the individuals with respect to that attribute. In the Population space the individuals provide the points, and the stimulus provides the attribute. These then are the respective functions of stimuli and individuals in a Population space. What kinds of behavior do psychologists observe in which these are the respective functions of the stimuli and the subjects? I would say that certain questionnaires, certain interest and neurotic inventories, and essay examinations represent the kinds of behavior that are typically mapped into Population spaces. There is a variety of possible methods of observing such behavior, but, speaking category-wise, the most typical are rating scales and free-answer protocols.

The rating scale method is the questionnaire item with a number of ordered alternatives—an example from a questionnaire used on soldiers during the war is the following:

Do you ever get so blue and discouraged that you wonder whether anything is worth while?

a) Hardly ever
b) Not so often
c) Pretty often
d) Very often

Such a procedure is formally equivalent to asking an individual to judge weights as being light, medium, or heavy. But one experiment is in a Population space, the other in a Stimulus space.

Free-answer protocols are illustrated by the essay examination, the open-ended questionnaire, and the interview. The individual is asked a question which in principle specifies an attribute, e.g., How do you feel about the farm policy? The individuals who are asked this question are playing the role of stimuli being evaluated with respect to an attribute. The protocols which emerge are analyzed for relations between the individuals, which constitute measures of them on this attribute. This is formally equivalent to the use of individual's evaluations of stimuli with respect to an attribute, which leads to information about where the stimuli are on the attribute selected by the individual.

One immediate consequence of this duality between Stimulus spaces and Population spaces is that any method for collecting or analyzing data constructed for either one of these spaces immediately becomes a potential method for collecting or analyzing data for the other. Thus we have, for example, Thurstone's Law of Comparative Judgment constructed for analyzing the judgment of individuals about stimuli and arriving at a Stimulus scale. Immediately there is implied the dual method of having stimuli make paired comparison judgments about individuals—as yet I have seen no good way of getting stimuli to do this. However, there are variations of this basic method of Thurstone's: the Method of Successive Intervals and the Method of Equal Appearing Intervals are used for constructing Stimulus spaces which do transfer completely to Population spaces. To transfer the Method of Successive Intervals to Population spaces you need to have stimuli sort people into piles. I suggest this is exactly what is done by those questionnaire items with multiple alternatives, e.g., from strongly agree to strongly disagree. Abstractly, we can look upon such behavior as stimuli sorting individuals into piles. In the Method of Successive Intervals the instructions to the subject are dualistically equivalent to the writing and editing of an item for a questionnaire or essay examination. I find it strikingly curious that we frequently tend to use five degrees of indorsement or five ordered steps in the alternatives, whereas in the Method of Successive Intervals we have individuals sort items into as many as eleven piles. Whether there is a profound reason behind this, or whether it is unjustified adherence to tradition, I am not sure.

Just as Thurstone's methods for Stimulus scales are transferable to Population spaces, so also are methods designed for the analysis of data in Population spaces alternative methods for Stimulus spaces, e.g., Lazarsfeld's methods of latent structure analysis could be used for scaling stimuli in Stimulus spaces by reversing the subscripts which identify stimuli and individuals, and obtaining the appropriate kind of judgments from individuals.

We see here, in fact, the relation of certain methodologies of the psychophysicist studying Stimulus spaces to those of the social psychologist studying Population spaces. What the first makes people do to the stimuli, the latter makes stimuli do to people. Their methods of doing research, collecting, and analyzing data are formally isomorphic with the roles of stimuli and people reversed. Surely, with reference to methodology, whatever one develops suggests a dual development for the other. Each delineates an attribute on which the objects of judgment are to be evaluated. In psychophysics the experimenter does this through his instructions to the subject; in questionnaires and essay examinations the experimenter does it through his careful writing of items. All the various experimental controls developed in one context, again, in principle, transfer to the other context with the reversal of roles between stimuli and people.

The greater status in measurement of psychophysics is due, at least in part, to the fact that an individual can compare two stimuli directly, whereas a stimulus cannot compare two individuals directly. We are much happier with the judgment of an individual as to which of two attitude statements he prefers to indorse than we are with the judgment of which of two individuals indorses a given statement more strongly. The reason for this is very simple. When we ask an individual which of two stimuli he prefers, we assume he has an implicit standard of measurement that is an ordered scale applicable to both stimuli. If we wish to compare two individuals as to which indorses a stimulus more strongly, we have to assume not only that they each have an implicit interval scale but also that the scales have the same origin and the same unit of measurement. Thus, if individual *A* says he would pay $10 for a picture and individual *B* $5, how do we know but that *A* has less value for money than *B*? Once we have made the assumption of an interpersonally comparable interval scale, there is no sense in reducing the data to a paired comparison—that would be throwing away information already assumed. This argument can be summed up by saying that the implicit standards of judgment of one person are presumably more stable over the two stimuli than the standards of two people over one stimulus. This might well be the consideration that underlies using fewer alternatives for an item in a questionnaire than the number of piles used in the Method of Successive Intervals.

When one looks at the differences between areas in this context one finds no justification for quarrels nor for differences in respectability. One area can use a system just as logically precise as the other, but the basic data observed in one area may be a weaker relation than is observed in the other area.

With Stimulus and Population spaces mutually described and related we turn briefly to Joint spaces. Here both stimuli and individuals come together, jointly specifying what the attributes will be; both have their own measures on these attributes. For example, consider an individual working an arithmetic problem. The arithmetic problem is represented by a point with measures on one or more components. The individual is similarly represented by a point with measures on these components. The response of the individual to the stimulus will be information about the relation of these two points in the Joint space.

By defining what this information is in different ways, one gets Guttman scalogram theory, test theory, one of Lazarsfeld's models, or my unfolding technique for the analysis of preferences. All of these are just different models for what a response on the phenotypic level means in terms of distances between pairs of points in the Joint genotypic space. [The relation of these various spaces to the classification of methodologies contained in (**3**) should be pointed out. The methods of collecting data which apply to Joint spaces are classified in Quadrants I and II, and the methods which apply in Stimulus spaces or in Population spaces are classified in Quadrants III and IV.]

In Joint spaces all the psychophysical methods for analyzing experiments concerned with thresholds, as distinct from those concerned with measuring only the stimulus magnitudes, are present. The individual in such experiments is regarded as having a threshold on an attribute, such as his sensitivity to light or his ability to discriminate pitch. This characteristic of the individual corresponds to a point in the genotypic space, which we have called his ideal on that attribute. The stimulus then is an increment of light or a difference between two tones, and the individual is asked whether he observes it. The stimulus is then also represented by a point in the space. The response of the individual is a formal relation on the pair of points in exactly the same manner as passing or failing an arithmetic item.

The data obtained from neurotic inventories and interest inventories are typically mapped into Joint spaces. For example, an individual is asked a question like "Are you shy?" which he is to answer *yes* or *no*. The individual is presumed to prossess and recognize his particular amount of shyness—this corresponds to a point in the genotypic space, which is his ideal. The question "Are you shy?" with the alternatives *yes* or *no* also corresponds to a point in the genotypic space which is that amount of shyness the individual feels he should have to say *yes*. This amount of shyness is formally equivalent to the difficulty of an arithmetic problem, the increment of light, or the difference between two tones in the preceding examples. Again the individual's response to the question is interpreted as a relation between the respective points.

That the data obtained from individual's indorsements or preferences between attitude statements may also be mapped into Joint spaces is too obvious to need further description. Most learning experiments are mapped into Joint spaces. A conditioning experiment, for example, is like an objective test given backwards: a combination of unconditioned stimulus and conditioning stimulus may be thought of as an item, and eliciting a conditioned response is equivalent to passing an item. Then the most difficult item in the test is presented first, i.e., the first presentation of conditioning stimulus and unconditioned stimulus, and the individual usually fails it. As learning takes place each successive presentation is essentially an easier item until items are so easy that the individual passes them all. It is interesting to note that the conditioning test has a different method of scoring from the objective mental test, e.g., the number of items taken to reach a certain number of items passed successively. One wonders why most objective tests should have a different convention. I do not object to different conventions, I just like to know what the logic behind them is.

So we have all these superficially different kinds of behavior: objective tests, certain psychophysical experiments, neurotic inventories, interest questionnaries, attitude scale studies, and conditioning experiments. All tend to develop their own methodologies and their own vocabulary—but all are formally isomorphic and hence their methodologies transferable from one to the other. A model for analysis of one of these kinds of data with a particular distance function, for example, immediately raises the question whether it does not also constitute a theory about each of the other seemingly different kinds of behavior. There are, of course, differences in the characteristics of the data one gets in these areas. In some areas experimentally independent replication is possible, in others not; in some areas the stability of a point associated with an individual or with a stimulus is greater or less than in other areas. But fundamentally these differences are quantitative, not qualitative, and the methodological contributions to any one area are in principle transferable to all the others.

Before going on to a psychological interpretation of Field spaces, I should digress for a moment and clear up a possible source of confusion. I have covered Stimulus, Population, and Joint spaces using repeated illustrations. There is a danger of certain misconceptions arising from the illustrations, which we must try to avoid. When I have illustrated one of these spaces, I have tried to follow the most conventional ways of analyzing such behavior, but the implication should *not* be drawn that the theory says there is only one kind of quantitative data or only one space into which any particular behavior can be mapped.

The act of a psychologist in putting his experimental data into one of these spaces (Joint, Stimulus, or Population) represents an optional decision on his part. The sense in which these decisions are optional is what I now want to make clear. The same behavior may be put into more than one of the spaces, thus reflecting different points of view or problems in the mind of the experimenter. The distinction between behavior and data, which was made earlier, is the relevant principle here. It is sometimes easy to see how the same behavior may be interpreted separately as two different kinds of data and consequently be put into different spaces. We have become so accustomed to certain conventions in the converting of manifest behavior into data that we sometimes neglect any mapping but the conventional one. While data is obtained from behavior by interpreting the behavior as a relation between points, it is up to the interpreter to decide what to identify as points and to define the properties of the relation.

Consider a study on nationality preferences. Let each subject make paired comparison judgments as to which nationality he prefers. In Thurstone's well-known study (**5**) such data were analyzed by the Law of Comparative Judgment and a scale obtained with the stimuli ranging on a one-dimensional continuum from most preferred to least preferred by the group as a whole. When the experimenter does this, he is regarding "preferability" as an attribute of stimuli and is saying that the individuals made task *B* judgments, substantive judgments, about the stimuli. The behavior is interpreted as an order relation on pairs of points, both of which are stimuli. He is saying the behavior belongs in a Stimulus space, and proceeds to construct a stimulus scale.

On the other hand, one could take the identical experiment and put it in a Joint space. In doing this, one would be assuming that the individuals were also points in the space and that their behavior is to be interpreted as an order relation on distances of the stimuli's points from the individual's point. Thus, the behavior is being put into a Joint space instead of a Stimulus space, and analysis of the data by multidimensional unfolding would yield a solution with both stimuli and individuals in the space.

There is nothing intrinsically correct about one of these procedures or wrong about the other. In the first instance, analyzing the data in a Stimulus space, one's problem is essentially that of amalgamating the preferences of individuals to arrive at a single preference scale, which in some sense best represents all the individuals. This is the problem of social utility or social choice. In the second instance, analyzing the data in a Joint space, one's problem is that of discovering the latent attributes underlying nationality preferences from which an individual's preferences could be derived. It might be parenthetically remarked that these two solutions would bear a certain interesting relation to each other, this relation has been developed in two previous publications (**2**, **4**).

Here we have taken an example of behavior and made the transition into two different kinds of data, analysis of which yields different results. We usually overlook this step that we take between behavior and data because this step, at least in some areas of research, is so conventional and immediate. Everyone can usually agree on what is or is not the right answer to an arithmetic problem; when an individual says this weight is heavier than that one, everyone usually agrees he means this weight is higher on an attribute of felt-heaviness than that one. But when we have an individual's answer to an essay examination question or his answer to an open-ended questionnaire item, we speak of "coding" them. This is the process of con-

verting the behavior to data by processing it through the mind of another
3 person to get statements of magnitude or relations—these data are what are
analyzed. It is important to note that what one analyzes is always data,
not behavior.

This distinction between behavior and data ~~now~~ becomes an even more important and relevant distinction as we turn to a ~~psychological~~ interpretation of Field spaces. To arrive at this interpretation we move along two dimensions of the Scale Grid simultaneously. In going from Joint to Stimulus spaces, the point associated with the individual became dependent on the stimuli being judged. There ceased to be a unique point characterizing the individual. Another way of looking at it is that the stimulus ceased to define the attribute with respect to which the judgments were made. In passing from Joint to Population spaces, the point associated with a stimulus became dependent on the individuals being judged. There ceased to be a unique point characterizing the stimulus. Another way of looking at this is that the individual ceased to define the attribute with respect to which the judgments were made.

In Joint spaces both stimuli and individuals are points and jointly define the attribute. In Stimulus spaces the stimuli are independent points, and the individuals are instructed to define the attribute. In Population spaces the individuals are independent points and the stimuli are instructed to define the attribute. Putting these together for Field spaces, we have the points for stimuli and individuals mutually dependent with neither instructed to define an attribute.

If you wanted to observe such behavior what would you do? You would present an individual with a stimulus that was of such an ambiguous nature it would not arouse any common attribute space in individuals. At the same time the individual would be totally uninstructed to respond with respect to any particular attribute space. This is my definition of what would be a perfect projective test situation. The behavior that is observed is associated with a point in a psychological space with which both the individual and the stimulus are identified.

It is to be noted that in all the other types of spaces (Joint, Stimulus, and Population) the attribute space is at least implicitly defined by the stimuli and/or by the instructions to the subject. Consider all the care given to selecting and wording items properly so that they will ask exactly the right question. This is nothing more than trying to limit the attribute space generating behavior. Exactly the same objective underlies the care in the communication of instructions to the observer in a psychophysical experiment. This care is taken to insure that he will ask the same question of every stimulus. It is then assumed that these precautions have succeeded, and the behavior is interpreted as information about a pair of points, or the distance between them, or a pair of distances. This mapping, done by definition, is what translates behavior into data.

In a Stimulus space, the individual was instructed to ask of the stimulus how heavy it was, or how esthetically pleasing it was, etc. An attribute was explicit, and so the behavior could be interpreted as magnitudes on an attribute and thus made into data. Analysis of such data leads to conclusions about *interstimulus differences.* In a Population space the roles are reversed: a stimulus comes to the individual and asks him how he feels about the farm policy. Again an attribute is explicit. The behavior is interpreted as magnitudes on an attribute, analysis of which leads to conclusions about *interindividual differences.*

In a Joint space both interpretations are possible because the behavior is interpreted as data on a relation between individuals and stimuli. In a Field space the point associated with the individual has merged with the point associated with the stimulus. The behavior is information about this point in a psychological space, a point in which the subject and the stimulus are inextricably identified.

The care given in Population spaces to selecting items for a questionnaire or essay examination in order to ask every individual the same question, and the care given in Stimulus spaces to phrasing instructions to the subject in order that he evaluate all the stimuli on the same attribute is now exercised in Field spaces so that precisely these effects will *not* occur. Every effort is made to insure a setting in which the stimulus will *not* suggest a particular attribute space, and every effort is made in the instructions to the subject *not* to suggest a particular attribute space. Herein lies both the strength and weakness of Field spaces. The protocol that emerges now constitutes a stimulus to be evaluated—so it is a stimulus to be located in a Stimulus space. In order to convert this protocol, this behavior, into data certain problems need to be solved. One is: what is the attribute or attributes which underlie the behavior? This is a new problem which had not previously arisen for any of the other spaces. Here now we have a protocol which is to be converted into a measure on some attributes. The first problem is: which attributes? This problem arises because there was not deliberately built into the stimulus nor into the individual constraints or instructions which would provide a simple answer.

It is immediately obvious that behavior in this area does not lead to interindividual comparisons, because there has been no instruction to the stimulus, no built-in device by virtue of which it can be assumed that a stimulus has evaluated each individual on the same attribute. If one person exhibits guilt feelings and another does not, one cannot conclude the latter person has less guilt feelings unless one can assume the stimulus was such as to make every individual reveal his guilt feelings. In this case, of course, such data could be put back in Population spaces, and interindividual comparisons would be possible.

It may help one to recognize and understand this problem if we point out that it is like having the answer of an individual to an essay examination question when you do not know what the question was. Consider, then, having the answers of several individuals, each to an unknown question; the problem is to decide which individual's answer represents more of something than another's. It seems to me the problem is meaningless if the question answered by an individual has been left up to him to select, hence each individual has perhaps selected a different question. One could say, well I can evaluate their relative command of English, or their vocabulary level, or their handwriting if it is a written protocol. This is entirely correct, of course, and amounts to saying this is the common attribute which the stimulus aroused in all individuals, hence they may legitimately be compared. This puts the behavior into a Population space, not a Field space. What I am talking about are those aspects of the behavior which are not attribute controlled and hence belong in a Field space.

Understandably enough, there have been instruments constructed, called projective instruments, which seek to avoid this particular problem. For instance, there are test instruments in which a picture suggests an attribute; the individual is asked to write a story which is presumed to reveal where he is on that attribute. Examples are Proshansky's Labor TAT, Johnson's Anglo-Spanish TAT, and toy play with negro and white dolls. If these instruments succeed in their purpose, then we have the stimulus coming to the individual with an attribute in mind and asking the individual where he is on it. These instruments then, if successful, are formally the same thing as a rather subtle essay examination or an interview by a laborer, a Mexican, or a Negro. The data that are obtained pertain to a Population space rather than a Field space, and interindividual comparisons are logically permissible. Such instruments, however, are not projective instruments in the sense of belonging in Field spaces. A further question then arises as to whether or not these instruments accomplish their purpose. If they fail to arouse the attribute which the experimenter subtly built into the stimulus then there is serious danger of drawing false conclusions.

Let us assume that the first problem is solved or can be solved—that we can look at the protocol and say what the attributes are which underlie the behavior. Then a second problem arises, which is the most fascinating and perhaps the most important of all: what does it mean that the individual selected these particular attributes to exhibit out of all of those possible in his repertoire? I think that no solution yet exists, but ultimately this problem must be answered in order to interpret projective instruments. This problem lies in the area of the psychology of the individual. Because the attributes were left up to the individual, their selection is a reflection of his internal dynamics. Because each individual is answering different questions, the behavior cannot be taken to reflect interindividual comparisons on a common attribute. On the contrary, and therein lies both its importance to psychology and its weakness as data, the behavior reflects *intra-individual* comparisons.

A protocol in a Field space reflects a point in a psychological space. When we know enough to interpret the protocol as a measure of that point in a known attribute space, then we shall be able to make comparisons between the points. I suspect that these will be comparisons on some hyperabstract attributes which will reflect intra-individual dynamics. The problems which must be solved to reach this stage are what I would regard as our ultimate measurement problems. Field spaces are a maximally significant domain of behavior. It is the area that reflects intra-individual differences to a degree that no other area does. There are fascinating and important problems for psychologists here. It is my thought and hope that the Scale Grid will help to delineate more clearly the basic measurement problems involved.

REFERENCES

1. Bentley, M. Early and late metric uses of the term "distance." *Am. J. Psychol.*, 1950, **63**, 619.
2. Coombs, C. H. A theory of psychological scaling. *Eng. Res. Inst. Bull. No.* 34, Ann Arbor, Mich.: Univ. Mich. Press, 1952.
3. Coombs, C. H. Theory and methods of social measurement. In L. Festinger and D. Katz (eds.), Research methods in the behaviorial sciences. New York: Dryden Press, 1953.
4. Coombs, C. H. Social choice and strength of preference. In R. M. Thrall, C. H. Coombs, and R. L. Davis (eds.), Decision processes. New York: Wiley, 1954.
5. Thurstone, L. L. An experimental study of nationality preferences. *J. Gen. Psychol.*, 1928, **1**, 405-424.

Manuscript received 9/4/56

PSYCHOMETRIKA—VOL. 21, NO. 4
DECEMBER, 1956

EFFICIENT ESTIMATION AND LOCAL IDENTIFICATION IN LATENT CLASS ANALYSIS*

RICHARD B. MCHUGH

IOWA STATE COLLEGE

Estimators which are efficient in the sense of having minimum asymptotic variance are obtained for the structural parameters of Lazarsfeld's latent class model of latent structure analysis. Sufficient conditions for the local identification of the structural parameters are also presented.

1. *Introduction*

A psychological test of p items or subtests will be denoted by $(x_1, x_2, \cdots, x_p)$. Data obtained from the administration of a psychological test can be processed in different ways. For example, in Lazarsfeld's latent structure analysis (**14**) the test data are treated qualitatively, i.e., as measurements on a nominal or at most an ordinal scale. On the other hand, in classical Spearman-Thurstone factor analysis test data are treated quantitatively, i.e., as measurements on an interval scale. Despite this difference there is a feature common to these two approaches, viz., the view that the αth examinee's observed test behavior $(x_{1\alpha}, x_{2\alpha}, \cdots, x_{p\alpha})$ may be thought of as being determined by his status $(\xi_{1\alpha}, \xi_{2\alpha}, \cdots, \xi_{\lambda\alpha})$ on λ traits or factors $(\xi_1, \xi_2, \cdots, \xi_\lambda)$ underlying the test. That is, observed test behavior (referred to as phenotypic or manifest) is viewed as only an indicant of, not as a direct measurement of, the underlying traits (referred to as genotypic or latent). [Among others, Coombs (**2**) and Stevens (**13**) offer justification for this approach to psychological measurement.]

Fundamental to this kind of test analysis are the concepts of *structure*, *model*, and *identification*. In section 3 below, the general formulation of these concepts due to Hurwicz (**5**), to Koopmans and Reiersøl (**6**), and to Reiersøl (**11**) is applied to a particular model of Lazarsfeld's latent structure analysis, viz., the latent class model.

In addition to (*a*) the pre-statistical problem of identification of the model, other problems which arise in this approach to the analysis of a psychological test include (*b*) the efficient estimation of the identifiable parameters of the model, and (*c*) the scaling problem, i.e., the assignment of trait measurements $(\xi_{1\alpha}, \xi_{2\alpha}, \cdots, \xi_{\lambda\alpha})$ to each examinee. Section 3 is con-

*The writer wishes to acknowledge with appreciation the helpful advice of Professors Leonid Hurwicz and Jacob Bearman of the University of Minnesota. Dr. John Gurland of Iowa State College assisted the author in clarifying certain points in this article.

cerned with problem (*a*) and section 4 with problem (*b*). [Problem (*c*) has been discussed for the latent class model by Lazarsfeld (**14**)]. Section 5 presents a statistical test of the goodness of fit of the structure, and in section 6 a numerical example is presented to illustrate the mechanics of computation.

Section 2 provides a resumé which makes certain features of Lazarsfeld's presentation (**14**) explicit in a form and notation convenient to the remainder of the discussion.

2. *The Latent Class Model*

Three distribution functions and an assumption concerning one of them constitute the essential ingredients of the latent class model.

(*i*) Let $G(x_1, x_2, \cdots, x_p)$ denote the distribution function of the p observable (manifest) random variables corresponding to the p items or subtests of the test. In practice, the subtests or items are usually dichotomous, i.e., the response of the αth subject to x_i, denoted by $x_{i\alpha}$, equals 1 for a *positive* response and 0 for a *negative* response ($i = 1, 2, \cdots, p$). Since the domain of the ith random variable, x_i, is 0 and 1, then $G(x_1, x_2, \cdots, x_p)$ is a p-dimensional binomial distribution function. Let the set of parameters defining the test response distribution, $G(x_1, x_2, \cdots, x_p)$, be denoted by $\{g_{i_1 i_2 \cdots i_p}\}$. These parameters are called *manifest marginal* probabilities. There are 2^p of the elements $g_{i_1 i_2 \cdots i_p}$, viz., $g_{12\cdots p}$, $g_{\bar{1}2\cdots p}$, $\cdots$, $g_{\bar{1}\bar{2}\cdots\bar{p}}$. By definition, $g_{12\cdots p} = P\{x_1 = 1, x_2 = 1, \cdots, x_p = 1\}$, $g_{\bar{1}2\cdots p} = P\{x_1 = 0, x_2 = 1, \cdots, x_p = 1\}$, and so forth to $g_{\bar{1}\bar{2}\cdots\bar{p}} = P\{x_1 = 0, x_2 = 0, \cdots, x_p = 0\}$.

In the latent model, then, the test may be interpreted as a random vector $(x_1, x_2, \cdots, x_i, \cdots, x_p)$ of p components—no effort is made to quantify the item response patterns. This may be contrasted with the mental test theory models, Gulliksen (**4**), Lord (**7**), where the test results are summarized in a total score, $\sum_{i=1}^{p} x_i$. The mental test theory models, therefore, use less of the information available from the test behavior than does the latent class model. In a given test situation, the loss in information may not be of practical significance if p is large, whereas the gain in computational convenience may be practically significant.

(*ii*) Let $F(\xi_1, \xi_2, \cdots, \xi_\lambda)$ denote the distribution function of the λ unobservable (latent) random variables. However, since the aim in practice (**14**) is generally to construct a *pure* test, i.e., a test measuring a single trait, it will be assumed here that $\lambda = 1$. That is, although the test in (*i*) is conceived as consisting of p variables, the trait ξ is postulated to be a one-dimensional random variable with distribution $F(\xi)$.

The domain of ξ need not be restricted to a point set of two elements, 0 and 1, as is the domain of x_i. In general, the domain of ξ consists of a point set of γ elements. These elements are termed *latent classes.* The generic class will be denoted by c ($c = 1, 2, \cdots, \gamma$). For example, in the numerical illustration in section 6, where the latent trait ξ is creativity in machine design, the case of $\gamma = 2$ is considered, viz., latent class $c = 1$ of creative machine designers and latent class $c = 2$ of non-creative machine designers. (By means of a chi square test of goodness of fit, however, as noted in section 6, a possible inference for this example is that $\gamma > 2$).

On the basis of this qualitative interpretation of the latent trait, $F(\xi)$ is a one-dimensional multinomial distribution function, specifically a one-dimensional γ-nomial. Let the set of parameters defining the latent trait distribution $F(\xi)$ be denoted by $\{f_c\}$. These parameters are termed *latent marginal* probabilities. There are γ of the elements f_c, viz., $f_1, f_2, \cdots, f_\gamma$. By definition, $f_c = P\{\xi = c\}$, $(c = 1, 2, \cdots, \gamma)$.

For this model, the conceptual measurement ξ_α of ξ for the αth examinee (i.e., ξ_α is conceived of as the value that would be obtained for the αth subject if the latent trait ξ could be measured directly) signifies an assignment of the αth subject to one of the latent classes $c = 1, 2, \cdots, \gamma$. This problem of scaling, i.e., of drawing a latent inference from the observed test response to the unobserved trait, is problem (*c*) cited in section 1.

(*iii*) The final distribution function of the latent class model is the distribution of the subtests $x_1, x_2, \cdots, x_p$ for fixed ξ. Let $G(x_1, x_2, \cdots, x_p \mid \xi = c)$ denote this conditional distribution function.

This conditional distribution is important because it makes explicit the relation between the observable variables x_i taken jointly and the unobservable ξ. For example, one of the 2^p parameters defining this p-dimensional binomial distribution $G(x_1, x_2, \cdots, x_p \mid \xi = c)$ is the probability of a conditional joint positive response, say $g_{12\cdots p|c}$. Defined for the αth subject, this is $g_{12\cdots p|c} = P\{x_{1\alpha} = 1, x_{2\alpha} = 1, \cdots, x_{p\alpha} = 1 \mid \xi = c\}$, indicating that the αth subject's test behavior depends probabilistically upon his status on the trait ξ.

Instead of dealing with the joint conditional distribution $G(x_1, x_2, \cdots, x_p \mid \xi = c)$, however, it may suffice to deal instead with the marginal conditional distributions $G(x_1 \mid \xi = c), G(x_2 \mid \xi = c), \cdots, G(x_p| \xi = c)$. This great simplification is possible if the psychological test has been constructed in such a way that

$$G(x_1, x_2, \cdots, x_p \mid \xi = c) \\ = G(x_1 \mid \xi = c)G(x_2 \mid \xi = c) \cdots G(x_p \mid \xi = c), \tag{2.1}$$

i.e., in case the subtests or items are jointly statistically independent for fixed trait level or latent class $\xi = c$. For a specified population of examinees, a test with this desirable property is called *pure* (**14**). That is, if (2.1) holds, each x_i involves a single common trait ξ. Without this factorability condition, the subtests or items would involve, ambiguously, several common traits $\xi_1, \xi_2, \cdots, \xi_\lambda$ $(1 < \lambda \leq p)$. In all that follows, Lazarsfeld's postulate (2.1) of a pure test will be assumed. The left and right sides of (2.1) will therefore be used interchangeably.

Because of their importance in the latent class model, the conditional distribution functions $G(x_i \mid \xi = c)$ of the various items, or subtests, are given the special title of *item trace functions* (**14**). These $G(x_i \mid \xi = c)$ are one-dimensional binomial distributions. By definition, the generic defining parameter is $g_{i;c} = P\{x_i = 1 \mid \xi = c\}$. The $g_{i;c}$ are referred to as *item conditional probabilities.*

To summarize, the latent class model involves the three distributions:

(*i*) $G(x_1, x_2, \cdots, x_p)$, defined by the set of 2^p parameters $\{g_{i_1 i_2 \cdots i_p}\}$;
(*ii*) $F(\xi)$, defined by the set of γ parameters $\{f_c\}$;
(*iii*) $G(x_1, x_2, \cdots, x_p \mid \xi = c)$, defined by the set of $p\gamma$ parameters

$$\{g_{i;c}\}, \qquad (i = 1, \cdots, p; c = 1, \cdots, \gamma).$$

TABLE 1

Schematic Representation of the Latent Class Model

One latent variable			ξ		
Domain of ξ (i.e.,, γ latent classes)		1	2...	c...	γ
Latent marginal probabilities		f_1	$f_2 \cdots$	$f_c \cdots$	f_γ
p manifest variables (items or subtests) x_i	Domain of x_i (i.e., dichotomy)	Item conditional probabilities			
x_1	1	$g_{1;1}$	$g_{1;2} \cdots$	$g_{1;c} \cdots$	$g_{1;\gamma}$
	0	$g_{\bar{1};1}$	$g_{\bar{1};2} \cdots$	$g_{\bar{1};c} \cdots$	$g_{\bar{1};\gamma}$
x_2	1	$g_{2;1}$	$g_{2;2} \cdots$	$g_{2;c} \cdots$	$g_{2;\gamma}$
	0	$g_{\bar{2};1}$	$g_{\bar{2};2} \cdots$	$g_{\bar{2};c} \cdots$	$g_{\bar{2};\gamma}$
$\vdots$	$\vdots$	$\vdots$			$\vdots$
x_i	1	$g_{i;1}$	$g_{i;2} \cdots$	$g_{i;c} \cdots$	$g_{i;\gamma}$
	0	$g_{\bar{i};1}$	$g_{\bar{i};2} \cdots$	$g_{\bar{i};c} \cdots$	$g_{\bar{i};\gamma}$
$\vdots$	$\vdots$	$\vdots$			$\vdots$
x_p	1	$g_{p;1}$	$g_{p;2} \cdots$	$g_{p;c} \cdots$	$g_{p;\gamma}$
	0	$g_{\bar{p};1}$	$g_{\bar{p};2} \cdots$	$g_{\bar{p};c} \cdots$	$g_{\bar{p};\gamma}$

Table 1 gives a convenient scheme for representing $F(\xi)$ and $G(x_1, x_2, \cdots, x_p \mid \xi = c)$, as defined by $\{f_c\}$ and $\{g_{i;c}\}$, respectively.

The three distributions (*i*), (*ii*), and (*iii*) are related by a set of 2^p equations for the latent class model:

$$\begin{aligned} g_{12\cdots p} &= \sum f_c g_{1;c} g_{2;c} \cdots g_{p;c}, \\ g_{\bar{1}2\cdots p} &= \sum f_c g_{\bar{1};c} g_{2;c} \cdots g_{p;c}, \\ &\cdots, \\ g_{\bar{1}\bar{2}\cdots\bar{p}} &= \sum f_c g_{\bar{1};c} g_{\bar{2};c} \cdots g_{\bar{p};c}, \end{aligned} \tag{2.2}$$

where the summation is from $c = 1$ to γ. The set (2.2), the accounting equations, follows from the fact that the distribution $G(x_1, x_2, \cdots, x_p)$ can be uniquely determined from the distributions $F(\xi)$ and $G(x_1, x_2, \cdots, x_p \mid \xi = c)$ by summing the product $F(\xi)G(x_1, x_2, \cdots, x_p \mid \xi = c)$ over the domain of ξ.

It should be noted that of the 2^p manifest marginal probabilities only $2^p - 1$ can be functionally independent since

$$\sum_D g_{i_1 i_2 \cdots i_p} = 1, \tag{2.3}$$

where the summation is over all the 2^p values of $i_1 \, i_2 \cdots i_p$, a domain which will be denoted by D. Similarly, of the latent marginal probabilities only $\gamma - 1$ can be independent because

$$\sum_{c=1}^{\gamma} f_c = 1.$$

Finally, the item conditional probabilities are related in general as

$$g_{i;c} + g_{\bar{i};c} = 1.$$

3. *Local Identification of the Structural Parameters f_c and $g_{i;c}$*

The parameters f_c and $g_{i;c}$ $(i = 1, \cdots, p; c = 1, \cdots, \gamma)$ of Table 1 are called *structural parameters* for the following reason. A structure, denoted here by $S^0 = \{F^0(\xi), G^0(x_1, x_2, \cdots, x_p \mid \xi = c)\}$, consists of the combination of particular, concretely specified distributions $F^0(\xi)$ and $G^0(x_1, x_2, \cdots, x_p \mid \xi = c)$. Equivalently, in terms of the defining parameters, $S^0 = \{f_c^0, g_{i;c}^0\}$, $(i = 1, \cdots, p; c = 1, \cdots, \gamma)$. That is, a structure is the combination of a set of particular numbers f_c^0 and $g_{i;c}^0$ for the parameters f_c and $g_{i;c}$. Table 4 presents an example of an estimated structure.

From (2.2) it is clear that a particular structure $\{f_c^0, g_{i;c}^0\}$ uniquely determines the set of parameters $\{g_{i_1 i_2 \cdots i_p}^0\}$ defining the distrubution function $G^0(x_1, x_2, \cdots, x_p)$ of the manifest subtests. The set $\{g_{i_1 i_2 \cdots i_p}^0\}$ is said to be generated by the structure S^0. Because each different (permissible) set of specific numbers f_c^0 and $g_{i;c}^0$ constitutes a different structure S^0, it is natural to consider the class $C = \{S\}$ of all such structures—this is the technical definition of a model. By definition, the *latent class model C* is the class of all structures $\{S\} = \{f_c, g_{i;c}\}$ conformable with the specifications of section 2. By contrast, an individual structure S^0 is a particular realization of the model, C.

It will be assumed in what follows that there is at least one structure. Then a problem of identification arises in a natural fashion. That is, a particular structure $\{f_c^0, g_{i;c}^0\}$ generates one and only one set of manifest parameters $\{g_{i_1 i_2 \cdots i_p}^0\}$. Hence the very practical question, called the *identification* question, is then posed, viz.: Does the converse proposition hold, i.e., can $\{g_{i_1 i_2 \cdots i_p}^0\}$ be generated by only the structure $\{f_c^0, g_{i;c}^0\}$? If so, then a given set of manifest marginal probabilities $\{g^0_{i_1 i_2 \cdots i_p}\}$ determines in principle one and only one structure $\{f^0_c , g^0_{i:c}\}$. If this were to occur, the latent class model $C = \{S\}$ would be said to identify (uniquely) the particular structure S^0, or the structure S^0 would be termed (uniquely) identifiable by the latent class model.

One necessary condition for identifiability of the structure S^0 or set of structural parameters f^0_c and $g^0_{i:c}$ is clearly that the number of independent pieces of manifest information be at least as great as the number of independent latent unknowns, i.e., the number, $2^p - 1$, of independent manifest parameters $g^0_{i_1 i_2 \cdots i_p}$ be at least as great as the number $\gamma - 1 + p\gamma$ of independent structural parameters f^0_c and $g^0_{i:c}$, or

$$2^p \geq \gamma(p + 1). \tag{3.1}$$

The answer to the general identification question may be that several structures besides S^0 generate the same manifest test distribution, i.e., the structure S^0 is not (uniquely) identifiable by the latent class model. However, a weaker form of identification, called *local identification*, may exist, viz., it may be that other structures generate $G^0(x_1 , x_2 , \cdots , x_p)$, but none of these lies in the neighborhood of S^0. (For convenience, the superscript "0" will be omitted in the following.)

THEOREM 1. *If specifications (i) through (v) hold for a structure of the latent class model, then the structural parameters f_c and $g_{i:c}$ are locally identifiable.*

(*i*) $2^p > \gamma - 1 + p\gamma$.

(*ii*) $\sum f_c g_{1:c} g_{2:c} \cdots g_{p:c} + \cdots + \sum f_c g_{\bar{1}:c} g_{\bar{2}:c} \cdots g_{\bar{p}:c} = 1.$

(*iii*) $\sum f_c g_{1:c} g_{2:c} \cdots g_{p:c} > 0,$

$\cdot \quad \cdot \quad \cdot \quad ,$

$\sum f_c g_{\bar{1}:c} g_{\bar{2}:c} \cdots g_{\bar{p}:c} > 0.$

(*iv*) $\sum f_c g_{1:c} g_{2:c} \cdots g_{p:c} ,$

$\cdot \quad \cdot \quad \cdot \quad ,$

$\sum f_c g_{\bar{1}:c} g_{\bar{2}:c} \cdots g_{\bar{p}:c} ,$

are continuous functions of $f_c , g_{1:c} , \cdots , g_{p:c}$ and possess continuous first and second partial derivatives.

(*v*) At least $\gamma - 1 + p\gamma$ of the expressions

$$\sum f_c g_{1:c} g_{2:c} \cdots g_{p:c} ,$$

$$\cdot \quad \cdot \quad \cdot \quad ,$$

$$\sum f_c g_{\bar{1}:c} g_{\bar{2}:c} \cdots g_{\bar{p}:c} ,$$

are functionally independent.

This theorem represents an application of a more general proposition, the heuristic origin for which is to be found in a classic paper by Fisher (**3**) on the limiting form of the Karl Pearson chi square criterion. A rigorous statement and proof of this more general proposition has been provided by Neyman (**8**, p. 250) in connection with the property of consistency of point estimators, i.e., the property of convergence in probability of an estimator to its corresponding parameter. That the proposition arises in connection with this property of estimators is not surprising since in ordinary practice only sample, not population, data are available. The following paragraph is concerned with the feasibility of the hypotheses (*i*) through (*v*) of Theorem 1 for the latent class model. This theorem is then related to the point estimation property of consistency.

Hypothesis (*i*) is equivalent to (3.1) and can be readily checked in any particular application. For example, it is satisfied in the example in section 6, where $p = 4$ and $\gamma = 2$, so that

$$2^p = 16 > 2 - 1 + (4)(2) = 9.$$

Hypothesis (*ii*) follows directly from (2.2) and (2.3). Specification (*iii*), on the other hand, cannot be verified directly; however, substantive considerations may strongly suggest it. For example, for $p = 4$ and $\gamma = 2$, if the contrary were true and

$$\sum_{c=1}^{2} f_c g_{1;c} g_{2;c} g_{3;c} g_{4;c} = 0,$$

then either $f_1 = 1$ (so $f_2 = 0$) and an even number of the $g_{i;2} = 0$, or an odd number of the $g_{i;2} = 0$—implications that are highly unlikely from the psychologist's viewpoint. (Thus, $f_1 = 1$ implies a degenerate application of the latent class model, viz., that almost all subjects belong in a single latent class—a vacuous scaling of ξ. Again, $g_{i;2} = 0$ implies that the ith subtest or item has almost no discriminating power and so is unlikely to be incorporated into the test initially.) The continuity hypothesis, specification (*iv*), clearly holds since the functions are polynomials in f_c and $g_{i;c}$. Finally, specification (*v*) is equivalent to asserting that the rank of the Jacobian matrix of the expressions is $\gamma - 1 + p\gamma$. It follows readily that full rank for the Jacobian is in turn equivalent to the requirement that the information matrix I (defined in section 6) be non-singular. This condition on I may be examined by direct algebra for a given model, or indirectly by numerical calculation of the approximate inverse of I, as in the example in section 6.

The connection between local identification of a structural parameter and the property of consistency is as follows: If it is possible to show that an estimator of the structural parameter θ, est. θ, is consistent, then θ must be locally identifiable. For example, in the latent class model if it is possible to show that the structural parameter f_c has an estimator est. f_c which converges stochastically to f_c, then f_c must be locally identifiable. This connection may be established by noting that if a structural parameter is not locally identifiable, i.e., in a neighborhood of the parameter, the parameter is not uniquely determined from the probability distribution of the observed variables, then every estimator of this parameter will fail to converge in probability to the parameter.

In terms of estimators, the application of the Neyman general proposition (**8**) to the latent class model means that Theorem 1 may be stated as follows: If the specifications (*i*) through (*v*) hold for a structure of the latent class model, then the estimators $\hat{f}_c$ and $\hat{g}_{i;c}$ of (4.2) are consistent, i.e., the set of solutions $\hat{f}_c$ and $\hat{g}_{i;c}$ of (4.2) are such that $\hat{f}_c$ converges in probability to f_c and $\hat{g}_{i;c}$ to $g_{i;c}$ as n tends to infinity.

4. *Efficient Estimation of the Structural Parameters f_c and $g_{i;c}$*

From the p-dimensional binomial distribution of test responses $G(x_1, x_2, \cdots, x_p)$ and from (2.2) it follows that the likelihood function of the structural parameters f_c and $g_{i;c}$, based on a sample of n examinees, is

$$L(f_c, g_{i;c}) = (\sum f_c g_{1;c} g_{2;c} \cdots g_{p;c})^{n_{12\cdots p}} (\sum f_c g_{\bar{1};c} g_{2;c} \cdots g_{p;c})^{n_{\bar{1}2\cdots p}} \\ \cdots (\sum f_c g_{\bar{1};c} g_{\bar{2};c} \cdots g_{\bar{p};c})^{n_{\bar{1}\bar{2}\cdots\bar{p}}}, \tag{4.1}$$

where $n_{i_1 i_2 \cdots i_p}$ is the number of subjects in the sample who give the response pattern $i_1 i_2 \cdots i_p$.

Application to this likelihood function of the method of maximum likelihood yields $\hat{f}_c$ and $\hat{g}_{i;c}$ as estimators of f_c and $g_{i;c}$. These estimators are the solutions of the set of equations obtained by differentiating the logarithm of (4.1), viz.:

$$\begin{aligned} &\text{(a)} \quad \frac{\partial(\log L)}{\partial f_c} = 0 \qquad (c = 1, \cdots, \gamma - 1), \\ &\text{(b)} \quad \frac{\partial(\log L)}{\partial g_{i;c}} = 0 \qquad (i = 1, \cdots, p; c = 1, \cdots, \gamma). \end{aligned} \tag{4.2}$$

Not all applications of the method of maximum likelihood result in estimators which possess the property of consistency [Neyman and Scott (**9**) give examples]. However, as noted in section 3, $\hat{f}_c$ and $\hat{g}_{i;c}$ are consistent if the specifications (*i*) through (*v*) of Theorem 1 obtain. Similarly, the routine of maximum likelihood by itself does not guarantee the property of efficiency for the resulting estimators [cf. (**9**), for examples]. In the latent class model, however, $\hat{f}_c$ and $\hat{g}_{i;c}$ are efficient if (*i*) through (*v*) obtain.

THEOREM 2. *If specifications (i)–(v) of Theorem 1 hold for a structure of the latent class model, then the estimators $\hat{f}_c$ and $\hat{g}_{i;c}$ are asymtotically efficient. That is, $\hat{f}_c$ and $\hat{g}_{i;c}$ have a joint asymptotic Gaussian distribution, and any other* *estimators of f_c and $g_{i;c}$ which are consistent and asymptotically Gaussian have asymptotic variances exceeding the asymptotic variances of $\hat{f}_c$ and $\hat{g}_{i;c}$.*

[Proof of the general proposition, of which Theorem 2 is a special case, can be found in Neyman (**8**, p. 250).]

As is the case in most applications of maximum likelihood, the solutions $\hat{f}_c$ and $\hat{g}_{i;c}$ to the likelihood equations (4.2) are not in general expressible in closed form. The indirect approach of iteration must therefore be employed. That is, a trial solution $\tilde{f}_c$ and $\tilde{g}_{i;c}$ must be assumed and a linear system solved for small, additive corrections $\delta\hat{f}_c$ and $\delta\hat{g}_{i;c}$, after which (6.1) is applied. A convenient mechanization of this procedure, due to R. A. Fisher, is called the *scoring system.* The technical details of the scoring system are given by Rao (**10**). A numerical example is given in section 6 to illustrate the computational procedure.

5. *A Statistical Test for the Number of Significant Latent Classes.*

The final problem of statistical inference considered here is that of formulating a statistical test of the hypothesis that $\gamma = \gamma_0$, where γ_0 denotes a specified number of latent classes.

A natural approach to this problem is direct examination of the goodness of fit of the manifest marginal probabilities $g_{i_1 i_2, \cdots, i_p}$, generated by the structure via (2.2), to the actual population manifest marginal parameters $g^*_{i_1 i_2, \cdots, i_p}$. Thus, if the discrepancy between the generated $G(x_1 , x_2 , \cdots , x_p)$ and the actual $G^*(x_1 , x_2 , \cdots , x_p)$ is substantial, the psychologist might postulate $\gamma = \gamma_0 + 1$ in an effort to improve the fit. In practice, however, direct examination is generally impossible; statistical inference is required since only the sample estimates $n_{i_1 i_2, \cdots, i_p}/n$ not the actual population parameters $g^*_{i_1 i_2, \cdots, i_p}$ are available. Here $n_{i_1 i_2, \cdots, i_p}$ is the sample observed response frequency corresponding to subtest response pattern $i_1 i_2 \cdots i_p$ (cf. Table 2).

For this formulation, the classical chi square goodness-of-fit test would be applicable if the structure were completely specified independently of the set of sample data $\{n_{i_1 i_2, \cdots, i_p}\}$. Thus, for a postulated γ_0 and known f_c and $g_{i;c}$, the parameters $g^*_{i_1 i_2, \cdots, i_p}$ of $G^*(x_1 , x_2 , \cdots , x_p)$ would be determined by (2.2). Hence the discrepancy between observed and theoretical frequencies, i.e., between the estimates $n_{i_1 i_2, \cdots, i_p}$ of $ng^*_{i_1 i_2, \cdots, i_p}$ and the $ng^*_{i_1 i_2, \cdots, i_p}$ as calculated from (2.2), could be tested by:

$$\chi^2 = \sum_D \{(n_{i_1 i_2, \cdots, i_p} - ng^*_{i_1 i_2, \cdots, i_p})^2/ng^*_{i_1 i_2, \cdots, i_p}\},$$

which is distributed approximately as chi square with degrees of freedom $2^p - 1$.

However, in practice the structural parameters f_c and $g_{i;c}$—and therefore $g^*_{i_1 i_2, \cdots, i_p}$ as obtained from (2.2)—must be estimated from the sample, hence the classical chi square test above is not valid without modification.

THEOREM 3. *If specifications (i)–(v) of Theorem 1 hold for a structure of the latent class model, then the following quantity in the limit as n tends to infinity is distributed as chi square:*

$$\chi^2 = \sum_D \{(n_{i_1 i_2 \cdots i_p} - \hat{n}_{i_1 i_2 \cdots i_p})^2 / \hat{n}_{i_1 i_2 \cdots i_p}\}, \tag{5.1}$$

with degrees of freedom $2^p - \gamma(p + 1)$, *i.e., degrees of freedom equal to the number of independent manifest parameters* $g_{i_1 i_2 \cdots i_p}$ *minus the number of structural parameters* f_c *and* $g_{i;c}$. Here $\hat{n}_{i_1 i_2 \cdots i_p}$ is the sample latent response frequency corresponding to subtest response pattern $i_1 i_2 \cdots i_p$; i.e., $\hat{n}_{i_1 i_2 \cdots i_p}$ is generated by the structure upon replacement of the structural parameters by the efficient estimates, e.g., for $p = 4$ and $\gamma = 2$ as in section 6,

$$\hat{n}_{1234} = n[\hat{f}_1 \hat{g}_{1;1} \hat{g}_{2;1} \hat{g}_{3;1} \hat{g}_{4;1} + \hat{f}_2 \hat{g}_{1;2} \hat{g}_{2;2} \hat{g}_{3;2} \hat{g}_{4;2}].$$

A rigorous generalization of this proposition is given by Neyman (**8**).

6. *Illustrative Example*

In order to illustrate the calculations needed, a latent class structure is estimated from data obtained by Schumacher, Maxson, and Martinek (**12**). Four machine design subtests, given to 137 engineers, were dichotomized into positive, 1, (above the subtest mean) and negative, 0, (below the subtest mean). Table 2, second column, gives the sample frequencies $n_{i_1 i_2 i_3 i_4}$ with which the observed response patterns $i_1 i_2 i_3 i_4$ occurred.

TABLE 2

Frequency of Occurrence of Response Patterns for the Four Machine Design Subtests (12)

Response patterns	Observed Frequencies	Generated Frequencies
$i_1 i_2 i_3 i_4$	$n_{i_1 i_2 i_3 i_4}$	$\hat{n}_{i_1 i_2 i_3 i_4}$
$1\ 2\ 3\ 4$	23	18.276211
$\bar{1}\ 2\ 3\ 4$	8	8.449612
$1\ \bar{2}\ 3\ 4$	6	6.882195
$1\ 2\ \bar{3}\ 4$	5	8.915412
$1\ 2\ 3\ \bar{4}$	5	8.789920
$\bar{1}\ \bar{2}\ 3\ 4$	9	3.676943
$\bar{1}\ 2\ \bar{3}\ 4$	3	5.386429
$\bar{1}\ 2\ 3\ \bar{4}$	2	4.545934
$1\ \bar{2}\ \bar{3}\ 4$	2	4.123700
$1\ \bar{2}\ 3\ \bar{4}$	3	3.602141
$1\ 2\ \bar{3}\ \bar{4}$	14	5.034202
$\bar{1}\ \bar{2}\ \bar{3}\ 4$	8	8.185339
$\bar{1}\ \bar{2}\ 3\ \bar{4}$	3	4.205489
$\bar{1}\ 2\ \bar{3}\ \bar{4}$	8	8.586886
$1\ \bar{2}\ \bar{3}\ \bar{4}$	4	5.689062
$\bar{1}\ \bar{2}\ \bar{3}\ \bar{4}$	34	32.650525

To apply the latent class model, some hypothesis concerning γ, the number of latent classes must be made. Subsequently, this hypothesis is tested to determine whether the number is in fact sufficient to account for the data. In this example, the hypothesis is: $\gamma = 2$, i.e., that two latent classes (*Creative* versus *Non-creative* on machine design) are sufficient to account for the observed response pattern frequencies.

To start the iterative scoring system of Fisher (**10**) it is necessary to find first approximations $\tilde{f}_c$ and $\tilde{g}_{i;c}$ to the structural parameters f_c and $g_{i;c}$. Then efficient estimators $\hat{f}_c$ and $\hat{g}_{i;c}$ are given to a first iteration by

$$\begin{aligned} \hat{f}_c &= \tilde{f}_c + \delta\hat{f}_c , \\ \hat{g}_{i;c} &= \tilde{g}_{i;c} + \delta\hat{g}_{i;c} , \end{aligned} \tag{6.1}$$

where $\delta\hat{f}_c$ and $\delta\hat{g}_{i;c}$ are the corrections to be added to the trial solutions. Moreover, one iteration of the scoring system is sufficient to yield estimators $\hat{f}_c$ and $\hat{g}_{i;c}$ which are fully efficient, provided the trial solutions $\tilde{f}_c$ and $\tilde{g}_{i;c}$ are consistent (**8**, p. 255). Since the Anderson-Lazarsfeld-Dudman estimators (**1**) have this property, they are used in the present example as the first approximation (Table 3) to the structure.

TABLE 3

First Approximation [Anderson, (1)] to an Efficient Estimate of the Latent Class Structure

Machine design ability	ξ	
Latent classes (domain of ξ)	1 ("creative")	2 ("non-creative")
Latent marginal probabilities	$\tilde{f}_1 = .516228$	$\tilde{f}_2 = .483772$

Subtests x_i	Domain of x_i	Subtest conditional probabilities	
x_1	1	$\tilde{g}_{1;1} = .729419$	$\tilde{g}_{1;2} =$ 080928
	0	$\tilde{g}_{\bar{1};1} = .270581$	$\tilde{g}_{\bar{1};2} = .919072$
x_2	1	$\tilde{g}_{2;1} = .745642$	$\tilde{g}_{2;2} = .188564$
	0	$\tilde{g}_{\bar{2};1} = .254358$	$\tilde{g}_{\bar{2};2} = .811436$
x_3	1	$\tilde{g}_{3;1} = .774096$	$\tilde{g}_{3;2} = .112673$
	0	$\tilde{g}_{\bar{3};1} = .225904$	$\tilde{g}_{\bar{3};2} = .887327$
x_4	1	$\tilde{g}_{4;1} = .741515$	$g_{4;2} = .226393$
	0	$\tilde{g}_{\bar{4};1} = .258485$	$\tilde{g}_{\bar{4};2} = .773607$

For the latent class structure postulated to underlie the data of Table 2, an efficient estimate is given in Table 4. As an example of Table 4, $\hat{f}_1 = .581917$ because from Table 3, $\tilde{f}_1 = .516228$ and, by the method of calculation explained below, $\delta\hat{f}_1 = .065689$; hence, using (6.1),

$$\hat{f}_1 = \tilde{f}_1 + \delta\hat{f}_1 = .516228 + .065689 = .581917.$$

TABLE 4

An Efficient Estimate of the Latent Class Structure

Machine design ability	ξ	
Latent classes (domain of ξ)	1 ("creative")	2("non-creative")
Latent marginal probabilities	$\hat{f}_1 = .581917$	$\hat{f}_2 = .418083$

Subtests x_i	Domain of x_i	Subtest conditional probabilities	
x_1	1	$\hat{g}_{1;1} = .686730$	$\hat{g}_{1;2} = .114614$
	0	$\hat{g}_{\bar{1};1} = .313270$	$\hat{g}_{\bar{1};2} = .885386$
x_2	1	$\hat{g}_{2;1} = .728419$	$\hat{g}_{2;2} = .173071$
	0	$\hat{g}_{\bar{2};1} = .271581$	$\hat{g}_{\bar{2};2} = .826929$
x_3	1	$\hat{g}_{3;1} = .676457$	$\hat{g}_{3;2} = .078556$
	0	$\hat{g}_{\bar{3};1} = .323543$	$\hat{g}_{\bar{3};2} = .921444$
x_4	1	$\hat{g}_{4;1} = .676904$	$\hat{g}_{4;2} = .173385$
	0	$\hat{g}_{\bar{4};1} = .323096$	$\hat{g}_{\bar{4};2} = .826615$

Finally, the corrections $\delta\hat{f}_c$ and $\delta\hat{g}_{i;c}$ are obtained by the scoring system from the relations:

$$\begin{aligned}
\delta\hat{f}_c &= \frac{1}{n}\sum_{c'=1}^{\gamma-1} n\tilde{I}^{(c)(c')}\tilde{S}_{c'} + \frac{1}{n}\sum_{i=1}^{p}\sum_{c''=1}^{\gamma} n\tilde{I}^{(c)(i;c'')}\tilde{S}_{i;c''} \\
&\qquad (c = 1, \cdots, \gamma - 1); \\
\delta\hat{g}_{i;c} &= \frac{1}{n}\sum_{c'=1}^{\gamma-1} n\tilde{I}^{(c')(i;c)}\tilde{S}_{c'} + \frac{1}{n}\sum_{i''=1}^{p}\sum_{c''=1}^{\gamma} n\tilde{I}^{(i;c)(i'';c'')}\tilde{S}_{i'';c''} \\
&\qquad (i = 1, \cdots, p; c = 1, \cdots, \gamma).
\end{aligned} \tag{6.2}$$

Here the tilde on the $\tilde{S}$ and the $\tilde{I}$ indicate evaluation at the trial solutions $\tilde{f}_c$ and $\tilde{g}_{i;c}$ of:

(*i*) the S functions, one for each structural parameter, which are defined as

$$\begin{aligned}
S_c &= \frac{\partial(\log L)}{\partial f_c} \qquad (c = 1, \cdots, \gamma - 1), \\
S_{i;c} &= \frac{\partial(\log L)}{\partial g_{i;c}} \qquad (i = 1, \cdots, p; c = 1, \cdots, \gamma),
\end{aligned} \tag{6.3}$$

i.e., simply a relabeling of the left side of (4.2). These functions are referred to as the *efficient scores* for the parameters f_c and $g_{i;c}$. For the data of Table 2, the efficient scores as evaluated at the first approximation are given in Table 5.

(*ii*) the inverse matrix of the I, or *information*, functions. Taking the

TABLE 5

Efficient Scores $\tilde{S}_c$ and $\tilde{S}_{i;c}$ Evaluated at the First Approximation

$\tilde{S}_1 = 10.635378$	
$\tilde{S}_{1;1} = 3.624402$	$\tilde{S}_{1;2} = 35.114865$
$\tilde{S}_{2;1} = 1.260259$	$\tilde{S}_{2;2} = 6.861212$
$\tilde{S}_{3;1} = -25.120099$	$\tilde{S}_{3;2} = -5.782032$
$\tilde{S}_{4;1} = -16.637863$	$\tilde{S}_{4;2} = -9.284267$

TABLE 6(a)

Information Matrix per Single Observation, Evaluated at the First Approximation

		f_1	$g_{1;1}$	...	$g_{3;2}$	$g_{4;2}$
1	f_1	3.048439	0.509498	...	0.678498	0.365129
2	$g_{1;1}$	0.509498	2.017739	...	-0.526254	-0.255081
3	$g_{2;1}$	0.354050	-0.124733	...	-0.392939	-0.178048
.	.	.	.		.	.
.	.	.	.		.	.
.	.	.	.		.	.
9	$g_{4;2}$	0.365129	-0.255081	...	-0.127786	2.136945

TABLE 6(b)

Inverse of Information Matrix per Single Observation, Evaluated at the First Approximation

		f_1	$g_{1;1}$	...	$g_{3;2}$	$g_{4;2}$
1	f_1	0.588484	0.254609	...	-0.211503	-0.183881
2	$g_{1;1}$	-0.254609	0.677139	...	0.202876	0.157569
3	$g_{2;1}$	-0.187074	0.118647	...	0.141973	0.109468
.	.	.	.		.	.
.	.	.	.		.	.
.	.	.	.		.	.
9	$g_{4;2}$	-0.183881	0.157569	...	0.093239	0.561545

parameters pairwise, the information functions, expressed per single observation, are:

$$
\begin{aligned}
\frac{1}{n} I_{(c)(c')} &= \frac{1}{n} E(S_c S_{c'}) \qquad (c, c' = 1, \cdots, \gamma - 1),\\
\frac{1}{n} I_{(i;c)(i'';c'')} &= \frac{1}{n} E(S_{i;c} S_{i'';c''})\\
&\qquad (i, i'' = 1, \cdots, p; c, c'' = 1, \cdots, \gamma),\\
\frac{1}{n} I_{(c')(i;c)} &= \frac{1}{n} E(S_{c'} S_{i;c})\\
&\qquad (c' = 1, \cdots, \gamma - 1; i = 1, \cdots, p; c = 1, \cdots, \gamma),
\end{aligned}
\tag{6.4}
$$

i.e., essentially the variances and covariances of the efficient scores. The square matrix of these elements, of order $\gamma - 1 + p\gamma$, is denoted by I and is called the *information* matrix. Hence in (6.2), $n\tilde{I}^{(c)(c')}$, $n\tilde{I}^{(c')(i;c)}$, and $n\tilde{I}^{(i;c)(i'';c'')}$ constitute the elements of the inverse of the information matrix, after evaluation at the trial solutions. For the present machine design data, the matrix $(1/n)\tilde{I}$ is given in Table 6(a) and $n\tilde{I}^{-1}$ in Table 6(b). Explicit formulas for the calculation of (6.3) and (6.4) are given in the Appendix.

As an example of the use of (6.2),

$$\delta\hat{f}_1 = \frac{1}{n}\,[n\tilde{I}^{(1)(1)}\tilde{S}_1 + n\tilde{I}^{(1)(1;1)}\tilde{S}_{1;1} + n\tilde{I}^{(1)(2;1)}\tilde{S}_{2;1} + \cdots + n\tilde{I}^{(1)(4;2)}\tilde{S}_{4;2}]$$

$$= \frac{1}{137}\,[(.588484)(10.635378) + (-.254609)(3.624402)$$

$$+ (-.187074)(1.260259) + \cdots + (-.183881)(-9.284267) = .065689,$$

using Table 5 and Table 6(b).

From (6.2) the crucial step in the scoring system is the inversion of the information matrix, i.e., the existence of I^{-1} is a necessary condition for identifiability. That the non-singular character of I is also a sufficient condition has been noted in section 3.

Since I^{-1} is the asymptotic variance-covariance matrix of the efficient estimators of the structural parameters, it is often desired to obtain the first approximation inverse matrix $\tilde{I}^{-1}$ explicitly. [As an example of the use of $\tilde{I}^{-1}$, from Table 6(b), the variance of $\hat{f}_1$ is approximately $\tilde{V}(\hat{f}_1) = .588484/137 = .00429$]. However, it is worth noting that if estimates of the large sample variances and covariances are not sought, then (6.2) is equivalent to solving a set of simultaneous linear equations in the corrections $\delta\hat{f}_c$ and $\delta\hat{g}_{i;c}$ without explicitly inverting the matrix $\tilde{I}$. That is, as a matrix equation (6.2) is $\hat{\delta} = \tilde{I}^{-1}\tilde{S}$, which is equivalent to $\tilde{I}\hat{\delta} = \tilde{S}$, where

$$\hat{\delta} = \begin{bmatrix} \delta\hat{f}_1 \\ \vdots \\ \delta\hat{g}_{p;\gamma} \end{bmatrix} \quad \text{and} \quad \tilde{S} = \begin{bmatrix} \tilde{S}_1 \\ \vdots \\ \tilde{S}_{p;\gamma} \end{bmatrix}.$$

For the machine design data, (5.1) is applied to Table 2 in order to test for the number of significant latent classes, yielding

$$\chi^2 = (23 - 18.276211)^2/18.276211 + \cdots$$

$$+ (34 - 32.650525)^2/32.650525 = 33.01$$

with $2^4 - 2(4 + 1) = 6$ degrees of freedom. Since the probability of a chi square this large by pure chance is less than .001, there may well be justification, with respect to the universe of examinees and psychological test used (**12**), for rejecting the hypothesis of $\gamma = 2$ latent classes for the latent class model and for assuming the existence of more than two classes.

Appendix

Evaluation of the $\tilde{S}$ and $\tilde{I}$ Functions for Tables 5 *and* 6(a).

(*i*) *The efficient scores $\tilde{S}$ of Table* 5.

Differentiation of the logarithm of (4.1) and evaluation at the first approximation leads to the general computation formulas

$$\begin{aligned} \tilde{S}_c &= \sum_D \tilde{q}_{(c)(i_1 i_2 \cdots i_p)} n_{i_1 i_2 \cdots i_p} , \\ \tilde{S}_{i;c} &= \sum_D \tilde{q}_{(i;c)(i_1 i_2 \cdots i_p)} n_{i_1 i_2 \cdots i_p} . \end{aligned} \tag{1}$$

Thus, the efficient scores are linear functions of the response frequencies $n_{i_1 i_2 \cdots i_p}$, with coefficients $\tilde{q}$ which are functions of $\tilde{f}_c$ and $\tilde{g}_{i;c}$, viz., the quotients:

$$\begin{aligned} \tilde{q}_{(c)(i_1 i_2 \cdots i_p)} &= \frac{c\text{th term of } (\tilde{g}_{i_1 i_2 \cdots i_p}/\tilde{f}_c)}{\tilde{g}_{i_1 i_2 \cdots i_p}} , \\ \tilde{q}_{(i;c)(i_1 i_2 \cdots i_p)} &= \frac{c\text{th term of } (\tilde{g}_{i_1 i_2 \cdots i_p}/\tilde{g}_{i;c})}{\tilde{g}_{i_1 i_2 \cdots i_p}} . \end{aligned} \tag{2}$$

That is, the denominator of a coefficient $\tilde{q}$ is that manifest marginal $g_{i_1 i_2 \cdots i_p}$ indicated by the second subscript on $\tilde{q}$, and evaluated at the trial solution. In order to obtain the numerator of $\tilde{q}$ readily, it is convenient in the determination of the denominator to compute each of the γ terms separately (later adding them, of course, to get the complete denominator). Then as (2) shows, the numerator is easily obtained in one operation, viz., by dividing the cth term of the denominator by $\tilde{f}_c$ (if dealing with $\tilde{S}_c$) or by $\tilde{g}_{i;c}$ (if dealing with $\tilde{S}_{i;c}$).

For example, from (1), for $p = 4$ and $\gamma = 2$,

$$\tilde{S}_{3;2} = \tilde{q}_{(3;2)(1234)} n_{1234} + \tilde{q}_{(3;2)(\bar{1}234)} n_{\bar{1}234} + \cdots + \tilde{q}_{(3;2)(\bar{1}\bar{2}\bar{3}\bar{4})} n_{\bar{1}\bar{2}\bar{3}\bar{4}} ,$$

which, for the data of section 6, is $= -5.782032$ as shown in Table 5. For the first coefficient, $\tilde{q}_{(3;2)(1234)}$, the manifest marginal for the denominator is g_{1234}, corresponding to the second subscript 1234 on $\tilde{q}$, which from (2.2) is

$$g_{1234} = f_1 g_{1;1} g_{2;1} g_{3;1} g_{4;1} + f_2 g_{1;2} g_{2;2} g_{3;2} g_{4;2} .$$

Hence, the denominator is obtained as the sum of the following two parts ($c = 1, 2 = \gamma$): $\tilde{f}_1 \tilde{g}_{1;1} \tilde{g}_{2;1} \tilde{g}_{3;1} \tilde{g}_{4;1}$ and $\tilde{f}_2 \tilde{g}_{1;2} \tilde{g}_{2;2} \tilde{g}_{3;2} \tilde{g}_{4;2}$. The numerator of $\tilde{q}_{(3;2)(1234)}$ is obtained by dividing $\tilde{g}_{3;2}$ into the $c =$ 2nd term of the denominator, $\tilde{f}_2\tilde{g}_{1;2}\tilde{g}_{2;2}\tilde{g}_{3;2}\tilde{g}_{4;2}$, giving $\tilde{f}_2\tilde{g}_{1;2}\tilde{g}_{2;2}\tilde{g}_{4;2}$. Hence by (2) the first coefficient is

$$\tilde{q}_{(3;2)(1234)} = \frac{\tilde{f}_2\tilde{g}_{1;2}\tilde{g}_{2;2}\tilde{g}_{4;2}}{\tilde{f}_1\tilde{g}_{1;1}\tilde{g}_{2;1}\tilde{g}_{3;1}\tilde{g}_{4;1} + \tilde{f}_2\tilde{g}_{1;2}\tilde{g}_{2;2}\tilde{g}_{3;2}\tilde{g}_{4;2}},$$

which can be calculated by direct substitution from Table 3 to be .010358. The other coefficients are obtained in a similar manner.

Actually, for a given efficient score $\tilde{S}$, only half of the numerators of the coefficients $\tilde{q}$ need be calculated. For example, out of 16 numerators for $\tilde{S}_{3;2}$ only 8 of the numerators need be computed, each of these being the negative of one of the remaining 8, e.g., $\tilde{f}_2\tilde{g}_{1;2}\tilde{g}_{2;2}\tilde{g}_{4;2}$, which occurs in $\tilde{q}_{(3;2)(1234)}$, is the negative of $-\tilde{f}_2\tilde{g}_{1;2}\tilde{g}_{2;2}\tilde{g}_{4;2}$, which occurs in $\tilde{q}_{(3;2)(12\bar{3}4)}$.

(*ii*) *The information functions, $\tilde{I}$, of Table* 6(a).

Evaluation of the expectations (6.4) at the first approximation leads to the general computational formulas:

$$\frac{1}{n}\tilde{I}_{(c)(c')} = \sum_D \frac{[\text{numerator of } \tilde{q}_{(c)(i_1i_2\cdots i_p)}][\text{numerator of } \tilde{q}_{(c')(i_1i_2\cdots i_p)}]}{\tilde{g}_{i_1i_2\cdots i_p}},$$

$$\frac{1}{n}\tilde{I}_{(i;c)(i';c')} = \sum_D \frac{[\text{numerator of } \tilde{q}_{(i;c)(i_1i_2\cdots i_p)}][\text{numerator of } \tilde{q}_{(i';c')(i_1i_2\cdots i_p)}]}{\tilde{g}_{i_1i_2\cdots i_p}}, \tag{3}$$

$$\frac{1}{n}\tilde{I}_{(c')(i;c)} = \sum_D \frac{[\text{numerator of } \tilde{q}_{(c')(i_1i_2\cdots i_p)}][\text{numerator of } \tilde{q}_{(i;c)(i_1i_2\cdots i_p)}]}{\tilde{g}_{i_1i_2\cdots i_p}}.$$

Thus, comparing (3) with (2), one notes that Table 6(a) can be obtained from the calculations already made in completing Table 5.

For example, from (3) and (2), for $p = 4$ and $\gamma = 2$,

$$\frac{1}{n}\tilde{I}_{(2;1)(3;2)} = \frac{[(\text{1st term of } \tilde{g}_{1234})/\tilde{g}_{2;1}][(\text{2nd term of } \tilde{g}_{1234})/\tilde{g}_{3;2}]}{\tilde{g}_{1234}}$$

$$+ \cdots + \frac{[(\text{1st term of } \tilde{g}_{\bar{1}\bar{2}\bar{3}\bar{4}})/\tilde{g}_{2;1}][(\text{2nd term of } \tilde{g}_{\bar{1}\bar{2}\bar{3}\bar{4}})/\tilde{g}_{3;2}]}{\tilde{g}_{\bar{1}\bar{2}\bar{3}\bar{4}}},$$

which for the data of section 6 is $-.392939$. Specifically, the first of these terms for $(1/n)\tilde{I}_{(2;1)(3;2)}$ is

$$\frac{[\tilde{f}_1\tilde{g}_{1;1}\tilde{g}_{3;1}\tilde{g}_{4;1}][\tilde{f}_2\tilde{g}_{1;2}\tilde{g}_{2;2}\tilde{g}_{4;2}]}{\tilde{f}_1\tilde{g}_{1;1}\tilde{g}_{2;1}\tilde{g}_{3;1}\tilde{g}_{4;1} + \tilde{f}_2\tilde{g}_{1;2}\tilde{g}_{2;2}\tilde{g}_{3;2}\tilde{g}_{4;2}},$$

which can be computed directly from the calculations previously made for $\tilde{q}_{(3;2)(1234)}$ of $\tilde{S}_{3;2}$ [as illustrated in (*i*) above] and for $\tilde{q}_{(2;1)(1234)}$ of $\tilde{S}_{2;1}$, which can be similarly illustrated.

REFERENCES

1. Anderson, T. W. On estimation of parameters in latent structure analysis. *Psychometrika*, 1954, **19**, 1-10.
2. Coombs, C. H. A theory of psychological scaling. Eng. Res. Bull. No. 34, University of Michigan, 1952.
3. Fisher, R. A. The conditions under which chi square measures the discrepancy between observation and hypothesis. *J. R. stat. Soc.*, 1924, **87**, 442-450.
4. Gulliksen, H. Theory of mental tests. New York; Wiley, 1950.
5. Hurwicz, L. Generalization of the concept of identification. Statistical inference in dynamic economic models. Cowles Commission Monograph 10. 1950, 245-257.
6. Koopmans, T. C. and Reiersøl, O. The identification of structural characteristics. *Ann. math. Stat.*, 1950, **21**, 165-181.
7. Lord, F. A theory of test scores. Psychometric Monograph, No. 7, 1952.
8. Neyman, J. Contribution to the theory of the χ^2-test. Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University California Press, 1949.
9. Neyman, J. and Scott, E. L. Consistent estimates based on partially consistent observations. *Econometrica*, 1948, **16**, 1-32.
10. Rao, C. R. Advanced statistical methods in biometric research. New York: Wiley, 1952.
11. Reiersøl, O. On the identifiability of parameters in Thurstone's multiple-factor analysis. *Psychometrika*, 1950, **15**, 121-149.
12. Schumacher, C. F., Maxson, G. R., and Martinek, H. Tests for creative ability in machine design. Annual report, 1953. Project ONR 458. Armed Services Technical Information Agency 21 284.
13. Stevens, S. S. Handbook of experimental psychology. New York: Wiley, 1951.
14. Stouffer, S. A., Lazarsfeld, P. F., et al. Measurement and prediction. Princeton: Princeton Univ. Press, 1950.

Manuscript received 5/12/55

Revised manuscript received 9/15/55

PSYCHOMETRIKA—VOL. 21, NO. 4
DECEMBER, 1956

THE SELECTION OF JUDGES FOR PREFERENCE TESTING*

R. DARRELL BOCK

UNIVERSITY OF CHICAGO

A scheme for choosing a few individuals whose preferences for given objects are most representative of those of a larger group of individuals is proposed. The method involves (*a*) quantifying the preferences of each individual so as to discriminate optimally among objects, (*b*) testing statistically whether or not a common preference continuum may be assumed for the quantified preferences, (*c*) constructing a linear estimator of values for the objects on this continuum, if it may be assumed, and (*d*) selecting as judges the least number of individuals whose quantified preferences, when used with this estimator, determine values for the objects with acceptable accuracy. A numerical example based on food preferences is presented.

I. *Introduction*

Preference studies, particularly of foods, frequently depend upon a limited number of judges who have been chosen from a larger group of individuals. Common practice is to choose those individuals whose repeated preferences for the same objects are most reliable. This assumes that all individuals reflect a common dimension of taste and that an individual whose preferences are reliable, but atypical, will not be encountered. A more thorough method would include a test for dimensionality and, if this condition is met, the selection of judges who are most representative of the group as a whole.

Let us assume that the preference of the ith individual for the jth object on the kth occasion is determined by a preference score g_{ijk}, which has the composition

$$g_{ijk} = \alpha_i\omega_j + e_{ijk} , \tag{1}$$

where α_i is a coefficient characterizing the ith individual, ω_j is a value characterizing the jth object, and e_{ijk} is a random error distributed as $N(0, \zeta_i^2)$ and independently of all α_i, ω_j, and other e_{ijk}. As usual, the variance of g_{ijk} over all occasions and objects is unity and the variance of ω_j over all objects is taken so that

$$\alpha_i^2 + \zeta_i^2 = 1. \tag{2}$$

*Preparation of this paper has been supported in part by the Quartermaster Food and Container Institute for the Armed Forces. Views or conclusions expressed herein are those of the author and do not necessarily reflect the views or indorsement of the Department of Defense.

The problem is to test the compatability of the observed preferences with (1) and to derive a scheme for estimating the ω_j which will most efficiently utilize a limited number of judges.

The preference scores are not observable, of course, but are presumed to underlie ratings, rankings, or paired comparisons of the objects, and are assumed to be estimable only by resort to a scaling method which quantifies the preferences. In the present context, an efficient and practical method is to choose scale values for the preferences which will yield a minimal estimate of ζ_i^2 . This type of scaling has been proposed by Fisher (**6**), in a different connection by Guttman (**7**, **8**), and by Maung (**16**), Johnson (**12**), and Bartlett (**4**). Related tests of significance have been discussed by Bartlett (**2**), Fisher (**5**, **6**), Williams (**19**), Marriott (**15**), and others. A suitable version of the method for the present purpose is as follows.

II. *Scaling the Preferences*

$$\begin{array}{cc|cccc|c}
 & & \multicolumn{4}{c}{\text{Objects}} & \\
 & & 1 & 2 & \cdots & q & \\
\hline
 & 1 & t_{11} & t_{12} & \cdots & t_{1q} & \\
 & 2 & t_{21} & t_{22} & \cdots & t_{2q} & \\
\text{Occasions} & \vdots & \vdots & \vdots & t_{kj} & \vdots & \\
 & p & t_{p1} & t_{p2} & \cdots & t_{pq} & \\
\hline\hline
 & 1 & n_{.11} & n_{.21} & \cdots & n_{.q1} & n_{..1} \\
 & 2 & n_{.12} & n_{.22} & \cdots & n_{.q2} & n_{..2} \\
t & \vdots & \vdots & \vdots & & \vdots & \vdots \\
 & r & n_{.1r} & n_{.2r} & \cdots & n_{.qr} & n_{..r} \\
\hline
 & & n_{.1.} & n_{.2.} & \cdots & n_{.q.} & n_{...}
\end{array} \tag{3}$$

t_{kj} designates a category of preference on the kth occasion for the jth object ($t = 1, 2, \cdots, r$):

(a) For ratings, t indicates a quasi-quantitative category such as good, fair, or poor.

(b) For rankings, t indicates a position in an ordering.

(c) For paired comparisons, t indicates the number of times one object is preferred to all others.

$n_{.jt}$ is the frequency with which the tth category occurs for the jth object over the p occasions.

$n_{..t}$ is the frequency of the tth category for all objects.

$$n_{.j.} = p,$$
$$n_{...} = qp = n,$$
$$r \leq q.$$

Representing the estimated score of the tth category by x_t, the average preference score for the jth object on p repetitions is defined by

$$g_{j.} = \frac{\sum_t x_t n_{.jt}}{\sum_t x_{.jt}} = \frac{\sum_t x_t n_{.jt}}{p}, \tag{4}$$

where a subscript for the ith individual is tacitly understood.

Since interest in preference testing is primarily in distinguishing quantitatively between objects, scores should be assigned to the categories which will discriminate optimally among the objects, where optimum discrimination is attained when the sum of squares between objects is a maximum with respect to that within objects. But for sums of squares:

$$\text{Between} + \text{Within} = \text{Total},$$

so this is equivalent to maximizing the between sum of squares subject to the condition that the total sum of squares remains constant. That is, assign values to x_t which maximize

$$\frac{1}{p} \sum_j \left(\sum_t x_t n_{.jt} \right)^2 - \frac{1}{n} \left(\sum_t x_t n_{..t} \right)^2, \tag{5}$$

subject to the condition that

$$\sum_t x_t^2 n_{..t} - \left(\sum_t x_t n_{..t} \right)^2 / n \tag{6}$$

be a finite constant.

Since this maximum is independent of the origin and scale of the x_t, impose the conditions that scores for the sample sum to zero and have unit variance. That is,

$$\sum_t x_t n_{..t} = 0, \tag{7}$$

and

$$\sum_t x_t^2 n_{..t} = n \quad . \tag{8}$$

Accordingly, (5) and (6) become

$$\sum_j \left(\sum_t x_t n_{.jt} \right)^2 / p, \tag{9}$$

and

$$\sum_t x_t^2 n_{..t} \,. \tag{10}$$

To express the required maximum explicitly as a function of the observed frequencies, it is convenient to adopt matrix notation and to define:

$x = [x_1 \; x_2 \; \cdots \; x_t \; \cdots \; x_r]$ = row vector of assigned scores;

$F = [n_{.jt}] = r \times q$ matrix of frequencies in (3);

$FF'/p = H$;

$D = [n_{..t}] = r \times r$ diagonal matrix of marginal frequencies in (3).

Then (9) becomes

$$\sum_j \Big(\sum_t x_t n_{.jt}\Big)^2/p = xHx', \tag{11}$$

and (10) becomes

$$\sum_t x_t^2 n_{..t} = xDx'. \tag{12}$$

The maximum of xHx', subject to the condition that $xDx' = n$, is obtained by differentiating and equating the resulting expression to zero:

$$xHx' - \eta^2(xDx' - n),$$

where η^2 is a Lagrangian multiplier. The result is

$$x(H - \eta^2 D) = 0, \tag{13}$$

which may be written

$$x(HD^{-1} - \eta^2 I) = 0. \tag{14}$$

For x to be non-null it is necessary that the determinant

$$| \, HD^{-1} - \eta^2 I \, | = 0, \tag{15}$$

where η^2 is a root of (15) and x is the corresponding latent vector of HD^{-1}.

An iterative method [cf. Hartree (**10**, p. 178ff.)] may be used to obtain latent vectors z_h of HD^{-1} with arbitrary origin and unit. The first of these vectors, with $\eta_1^2 = 1$ and $z_1 = (1 \; 1 \; \cdots \; 1)$, is trivial and can be avoided by replacing F with the corresponding matrix of deviations from expectation. The second latent vector provides the required maximum of $xHD^{-1}x'$, and its associated root η_2^2 is, as the notation suggests, a correlation ratio giving the fraction of sum of squares between objects. For, noting that a linear transformation of a latent vector is a latent vector, from (13)

$$z_2 H - \eta_2^2 z_2 D = 0.$$

Transposing and post-multiplying by z_2' ,

$$z_2 H z_2' = \eta_2^2 z_2 D z_2' = \eta_2^2 n. \tag{16}$$

Similarly $\hat{\zeta}^2$, which estimates ζ^2, is the fraction of sum of squares within objects, and

$$\eta_2^2 + \hat{\zeta}^2 = 1. \tag{17}$$

To transform z_2 into x, impose conditions (7) and (8) to obtain

$$x_t = u(z_t - v), \tag{18}$$

where

$$v = \sum_t z_t n_{..t}/n, \tag{19}$$

and

$$u = \frac{\sqrt{n}}{\sqrt{\sum_t (z - v)^2 n_{..t}}}. \tag{20}$$

In general, HD^{-1} has $r - 1$ non-trivial latent vectors, of which the first, x_2 , provides estimates of the maximally discriminating scores for the preference categories. In the present context further interpretation for any remaining vector, even if its associated root could be shown significant, does not seem necessary, although it should be noted that relevant interpretations are available (**9**).

The statistical significance of the scores may be questioned on two counts: If they are to be useful for discriminating among the objects, variance attributable to differences in column means (mean preference scores) must be significant. Fisher (**6**) suggests that an approximate test, based on an analysis of variance, be made directly from the value of η_2^2 as shown in Table 1.

TABLE 1
Form of Analysis of Variance for Derived Scores

Source of variation	Degrees of freedom	Sums of squares	Mean squares	F
Between objects	$q-1+r-1$	$\eta_2^2 n$	$\dfrac{\eta_2^2 n}{q+r-2}$	$\dfrac{\eta_2^2 q(p-1)-(r-1)}{(1-\eta_2^2)(q+r-2)}$
Residual	$q(p-1)-(r-1)$	$n-\eta_2^2 n$	$\dfrac{n(1-\eta_2^2)}{q(p-1)-(r-1)}$	
Total	$n-1$	n		

[Cf. Williams (19).] The degrees of freedom are adjusted for the arbitrary constants fitted by adding $r - 1$ to those between objects and subtracting $r - 1$ from those of the residual. The scores may be considered useful only if F is significant.

It is also of interest to test whether these derived, maximally discriminating scores are any real improvement on arbitrarily assigned scores, e.g., $1, 2, 3, \cdots, r$. Again following Fisher (6), this may be done by testing the significance of variance between objects, computed from the derived scores, which remains after variance attributable to the assigned scores has been removed. The appropriate analysis of covariance is shown in Table 2. (The vector of arbitrarily assigned scores, must satisfy the relation $\sum_t \xi_t n_{..t} = 0$.)

TABLE 2
Form of Analysis of Covariance of Assigned and Derived Scores

Source of variation	Sums of squares (assigned scores)	Sums of cross-products	Sums of squares (derived scores)
Between objects	$\xi H \xi'$	$\xi H x'$	$x H x'$
Residual	$\xi D \xi' - \xi H \xi'$	$\xi D x' - \xi H x'$	$x D x' - x H x'$
Total	$\xi D \xi'$	$\xi D x'$	$x D x'$

The analysis of variance of the derived scores, eliminating the assigned scores, is shown in Table 3. The degrees of freedom between objects has been reduced by 1, since the elimination of ξ leaves only $r - 2$ scores adjustable.

TABLE 3
Form of Analysis of Variance of Derived Scores, Eliminating Assigned Scores

Source of variation	Degrees of freedom	Sums of squares	Mean squares	F
Between objects	$q+r-3$	SS_b : Obtained by subtraction	$\dfrac{\mathrm{SS}_b}{q+r-3}$	$\dfrac{\mathrm{SS}_b q(p-1)-(r-1)}{\mathrm{SS}_r(q+r-3)}$
Residual	$q(p-1) -(r-1)$	$\mathrm{SS}_r = \{xDx' - xHx'\} - \left\{\dfrac{(\xi Dx' - \xi Hx')^2}{\xi D\xi' - \xi H\xi'}\right\}$	$\dfrac{\mathrm{SS}_r}{q(p-1)-(r-1)}$	
Total	$n-2$	$\mathrm{SS}_t = xDx' - \dfrac{(\xi Dx')^2}{\xi D\xi'}$		

A significant value of F is necessary, of course, if the derived scores are to be preferred to more convenient assigned scores. As before, this test is only approximate. For a more exact treatment, see Bartlett (4).

III. *Estimating the* ω_j

If written explicitly, the scaled preferences of N individuals for g objects on p occasions would appear as in Table 4. For subsequent computations only the mean preference scores are needed and may be obtained by the relation

TABLE 4

Preference Scores and Mean Preference Scores

	Objects												
	1				2				⋯	q			
	Individuals				Individuals				⋯	Individuals			
Occasions	1	2	⋯	N	1	2	⋯	N	⋯	1	2	⋯	N
1	g_{111}	g_{211}	⋯	g_{N11}	g_{121}	g_{221}	⋯	g_{N21}	⋯	g_{1q1}	g_{2q1}	⋯	g_{Nq1}
2	g_{112}	g_{212}	⋯	g_{N12}	g_{122}	g_{222}	⋯	g_{N22}	⋯	g_{1q2}	g_{2q2}	⋯	g_{Nq2}
⋮	⋮	⋮		⋮	⋮	⋮		⋮		⋮	⋮		⋮
p	g_{11p}	g_{21p}	⋯	g_{N1p}	g_{12p}	g_{22p}	⋯	g_{N2p}	⋯	g_{1qp}	g_{2qp}	⋯	g_{Nqp}
Means	$g_{11.}$	$g_{21.}$	⋯	$g_{N1.}$	$g_{12.}$	$g_{22.}$	⋯	$g_{N2.}$	⋯	$g_{1q.}$	$g_{2q.}$	⋯	$g_{Nq.}$

$$g_{ij.} = x_i n_{ij}/p, \tag{21}$$

where x_i is the first non-trivial latent vector for the ith individual under the transformation (18), and n_{ij} is the jth column of the corresponding F matrix.

A method of estimating the ω_j from the preference scores of the N individuals is needed. Resorting again to the requirement of optimal discrimination among the objects, choose as an estimator a linear compound of the preference scores which maximizes the sum of squares of the combined scores between objects on condition that the total sum of squares of the combined scores is constant. Let the estimate of ω_j on the kth occasion be

$$w_{jk} = \sum_i y_i g_{ijk} , \tag{22}$$

and for all occasions

$$w_{j.} = \sum_i y_i g_{ij.} . \tag{23}$$

The analysis of variance of the preference scores with respect to this linear compound may be represented as:

	SS	DF
Between Objects	yMy'	$q - 1$
Within Objects	yBy'	$q(p - 1)$
Total	yTy'	$qp - 1$

where M, B, and T are, respectively, matrices of between objects, within objects, and total sums of squares and cross products for the preference scores as represented in Table 4. Only M must be computed explicitly, with

$$M = pGG', \tag{24}$$

where G is an $N \times q$ matrix of mean preference scores $[g_{ij.}]$.

The coefficients of the required optimally discriminating linear compound are the elements of y determined from

$$y(M - \lambda T) = 0, \tag{25}$$

where λ is the largest root of

$$| M - \lambda T | = 0. \tag{26}$$

If the preferences are made independently, and the objects are identical on successive occasions, the assumption that the preferences of different individuals for the same object are uncorrelated over occasions appears justified. In this case the off-diagonal elements of T and M are asymptotically equal and

$$B = T - M$$

approximates an $N \times N$ diagonal matrix with elements

$$b_{ii} = \sum_j \sum_k g_{ijk}^2 - p \sum_j g_{ij.}^2 \ .$$

By (8) and (16) this becomes

$$b_{ii} = n - (px_iF_iF_i'x_i')/p^2$$
$$= n - n\eta_i^2 \ ,$$

and by (17)

$$b_{ii} = n\zeta_i^2 \ . \tag{27}$$

Taking $Z = [n\hat{\zeta}_i^2]$ as an approximation of B,

$$| M - \lambda(M + Z) | = 0,$$

or, letting

$$\mu = m\lambda/(1 - \lambda), \tag{28}$$

with $m = q(p - 1) - (r - 1)$, the additional $r - 1$ degrees of freedom being removed because of the constants fitted in scaling,

$$| MZ^{-1} - \mu I/m | = 0. \tag{29}$$

Taking the largest root of (29), y may be determined from

$$y(MZ^{-1} - \mu I/m) = 0. \tag{30}$$

[Cf. Bartlett (**3**).]

If μ_1 is clearly significant, a basis for testing of the significance of variation attributable to the remaining roots is provided by Bartlett's (**2**) approximation for Wilks' ratio

$$\chi^2 = -[n - \tfrac{1}{2}(N + q)] \sum_{s=1}^{\min(N,q-1)} \log_e (1 - \mu_s). \tag{31}$$

For large n this becomes approximately

$$\chi^2 = \sum_{s=1}^{\min(N,q-1)} \mu_s \, . \tag{32}$$

That is, the latent vectors associated with each of the canonical variances transform MZ^{-1} into $\min(N, q - 1)$ asymptotically independent quadratic forms, each distributed as χ^2 with degrees of freedom equal to the number of arbitrary constants fitted. Specifically, the sum of the canonical variances,

$$m \cdot \mathrm{tr}\,(MZ^{-1}) = \mu_1 + \mu_2 + \cdots + \mu_{\min(N,q-1)} \, , \tag{33}$$

is distributed as the sum of $\min(N, q - 1)$ asymptotically independent chi squares with degrees of freedom

$$N(q - 1) = (N + q - 2) + (N + q - 4) + \cdots \\ + \{N - q - 2[\min (N, q - 1)]\}.$$

To test the residual variance, eliminate that of the first root by taking

$$\chi^2 = m \cdot \mathrm{tr}\,(MZ^{-1}) - \mu_1 \tag{34}$$

on $N(q - 1) - (N + q - 2)$ degrees of freedom.

Equation (29) has in general $\min(N, q - 1)$ roots, $\mu_1 , \mu_2 , \cdots , \mu_s$, when ordered by size, giving the canonical variances of MZ^{-1}. If (1) holds, only μ_1 will be significant and the linear compound specified by the corresponding latent vector y_1 will account for all the significant variation in the preference scores. In critical cases (31) would be somewhat more accurate.

It is of interest to note that if (1) holds and p is large,

$$g_{ijs} = \alpha_i \omega_j$$

approximately, and

$$G = \alpha'\omega.$$

If

$$\omega\omega'/q = 1,$$

$$GG'/q = \alpha'\alpha = M/n.$$

Then (30) becomes approximately

$$y(n\alpha'\alpha - \mu B/n) = 0,$$

where

$$| \, n\alpha'\alpha - \mu B/n \, | = 0$$

has only one root:

$$\mu = n^2 \alpha B^{-1} \alpha', \tag{35}$$

and one associated latent vector:

$$\begin{aligned} y &= n\alpha B^{-1} \\ &\cong n\alpha Z^{-1}, \end{aligned} \tag{36}$$

a result identical in form with that obtained when constructing a linear discriminant function for two groups [cf. Kendall (**14**, vol. II, p. 341)] and similar also to that for the estimation of mental factors [cf. Holzinger and Harman (**11**, p. 322)]. Furthermore, from (2) and (17)

$$\alpha_i^2 = \eta_i^2 ,$$

and, asymptotically,

$$Z = [n\zeta_i^2],$$

or, by (36)

$$y_i = \eta_i/(1 - \eta_i^2). \tag{37}$$

Hence, when unidimensionality of tastes can be assumed, the optimal weights for combining preferences are strictly analogous to those for combining tests scores on the basis of reliability [cf. Kelley (**13**, p. 211)]. For, if α is considered the correlation coefficient between a fallible score and a theoretical "true" score, then

$$\alpha_i^2 = r_{i\infty} r_{i\infty} = r_{ii} = \eta_i^2 ,$$

and

$$y_i = \sqrt{r_{ii}}/(1 - r_{ii}).$$

IV. *Selecting the Judges*

In the sample the best estimator of ω_i is given by the first latent vector y associated with the solution of (30). Individuals for whom the corresponding elements of y are larger are to be preferred as judges. If k such individuals are selected, the question arises whether the loss of information caused by excluding the preferences of the remaining $N - k$ individuals from the estimates of the ω_j is appreciable.

If y_k is the vector y after the elements for the unwanted individuals have been dropped, and M_k and Z_k^{-1} are M and Z^{-1} with corresponding $N - k$ rows and columns dropped, then adjusting y_k so that $y_k y_k' = 1$, the portion of the canonical variance μ_1 which is attributable to the k selected individuals is

$$v_k = m y_k M_k Z_k^{-1} y_k' \, . \tag{38}$$

The significance of the remaining variance can be tested approximately by taking

$$\mu_1 - v_k \tag{39}$$

as χ^2 with $(N - k)(q - 1)$ degrees of freedom [cf. Rao (**18**, p. 257)]. If this variance is insignificant, or appropriately small with respect to μ_1 , the preferences of the k selected judges can be considered adequate for estimating the ω_j .

Since the preferences within objects are uncorrelated, the elements remaining in y_k are unchanged except in scale. If it is convenient to have standarized values for the w_{jk} within objects and between occasions, the elements of y_k should be multiplied by

$$\left[(\hat{\zeta}_1^2 y_1^2 + \hat{\zeta}_2^2 y_2^2 + \cdots + \hat{\zeta}_k^2 y_k^2) \Big/ m\right]^{-1/2} . \tag{40}$$

V. *Related Problems*

It should be understood that the sample dealt with here is, in effect, a series of occasions out of infinitely many in which the preferences of the same individuals might be repeated under the same conditions. Predictions made on the basis of this sample apply only to future preferences of these individuals and are not immediately generalizable to the population from which the individuals are drawn. When a single significant dimension of taste is found, however, this limitation is less serious because there is no evidence which contradicts the assumption that a new group from the population will share the same dimension of taste and show the same relative preferences for the objects. When significant additional dimensions are found, the problem becomes much less tractable but in some ways more interesting. Additional dimensions raise the question of what and how many attributes in the objects are being evaluated by the individuals and reflected in the preferences. This is essentially a problem in multidimensional psychophysics; the type of canonical analysis used in the present study appears to be an alternative to approaching the problem by a direct reduction of a table of proportions from paired comparisons of the objects. Preference scores for the additional dimensions may be estimated using the other significant latent vectors from (30), and rotation of these vectors to yield more meaningful scores is permissible. In practical applications, determining dimensionality could be important because it would indicate the minimum number of attributes which must be controlled if the objects are to be produced in uniform quality. A natural extension of this would be to identify the attributes physically or chemically and subject them to planned control and development.

The converse problem would be to identify and characterize by independent variables those individuals whose preferences are responsible for or most representative of the various dimensions of taste. If individuals with atypical tastes were few, they could be excluded from further preference testing and the remaining homogeneous group considered representative of the ultimate consumers of the objects. Alternatively, it might be found that distinct subgroups of the individuals with different tastes can be identified in terms of sex, race, income-group, etc. A set of separate panels representing each of these subgroups would, so to speak, "span the space of preferences" in the population. If the proportion of these groups in the population were known, the reaction of the total population to new objects could be predicted by a weighted combination of the preferences of the subgroups. Finally, the relationships among the preferences of the groups could be studied by a canonical analysis or a multivariate analysis of variance in which the "individuals" of this study are replaced by groups and "occasions" by individuals within groups. A problem of this sort based on a national food preference survey is now under study.

VI. *Numerical Example*

Oberman and Li (**17**) report ten replicate ratings (p) of ten individuals (N) for pastries baked from five different fats (q). Ratings were made on a five-point scale (r): 1 = not edible; 2 = poor; 3 = fair; 4 = good; 5 = excellent. For the first individual, the ratings of the five pastries are shown in Table 5.

The body of Table 5 is the matrix F; the frequencies on the right are the elements of the diagonal matrix D. Then:

$$pH = FF' = \begin{bmatrix} 62 & 42 & 35 & 01 & 00 \\ 42 & 60 & 34 & 04 & 00 \\ 35 & 34 & 38 & 10 & 03 \\ 01 & 04 & 10 & 37 & 18 \\ 00 & 00 & 03 & 18 & 09 \end{bmatrix}$$

$$\text{Column sum:} \quad 140 \quad 140 \quad 120 \quad 70 \quad 30$$

The rows and columns of pH must sum to $pn_{..t}$. To form HD^{-1}, divide elements in the tth column of pH by $pn_{..t}$. It is convenient at the same time to remove the expectation by subtracting $1/r$ from each of the resulting elements. Designate the resulting matrix by T:

$$T = \begin{bmatrix} .2429 & .1000 & .0917 & -.1857 & -.2000 \\ .1000 & .2286 & .0833 & -.1429 & -.2000 \\ .0500 & .0428 & .1167 & -.0571 & -.1000 \\ -.1929 & -.1714 & -.1167 & .3286 & .4000 \\ -.2000 & -.2000 & -.1750 & .0571 & .1000 \end{bmatrix}$$

$$\text{Column sum:} \quad .0000 \quad .0000 \quad .0000 \quad .0000 \quad .0000$$

Columns of T must sum to zero. [Note: χ^2 for F, with $(q - 1)$ $(r - 1)$ degrees of freedom is $n(\text{tr } T)$.]

The first latent vector of T is z_2, the first non-trivial latent vector of HD^{-1}. To extract z_2 from T, premultiply by a trial vector orthogonal to the trivial vector (1 1 1 1 1). The vector (1 1 0 -1 -1) is convenient. For these data, five iterations produce the following essentially accurate estimate:

$$z_2 \cong (.8325 \quad .7831 \quad .6021 \quad -.7721 \quad -1.000),$$

with associated root $\eta_2^2 = .7921$. Using (18), (19), and (20), impose conditions (7) and (8) to obtain

$$(.6292 \quad .5514 \quad .2694 \quad -1.8691 \quad -2.2260).$$

Analysis of variance based on η_2^2 shows significant discrimination between fats (Table 6). Analysis of covariance of assigned scores 1, 2, 3, 4, 5 [adjusted to $\xi = (1.42\ 0.42\ -.58\ -1.58\ -2.58)$ so that $\sum_t \xi_t n_{..t} = 0$] and the derived scores is shown in Table 7. Analysis of variance of the derived scores, eliminating the assigned scores, shows that the derived scores discriminate between fats significantly better than the assigned scores (Table 8).

Scales for the remaining individuals are constructed in the same way. The correlation ratios and results of the significance tests for each individual are shown in Table 9. The mean preference scores of each individual for the pastries are shown in Table 10, the body of which comprises the matrix G. (Rows of G must sum to zero.) The elements of the diagonal matrix Z computed by (27) and (17) from the correlation ratios of Table 9 are:

10.395, 32.690, 32.345, 34.885, 19.355, 13.795, 31.625, 3.365, 24.845, 9.265.

The product $MZ^{-1} = pGG'Z^{-1}$ is given in Table 11.

The first latent vector of MZ^{-1} is extracted by premultiplying by $(1 \cdots 1)$

TABLE 5

Preferences of the First Individual

Categories (t)	Fats					Total
	1	2	3	4	5	
5	4	3	6	1	0	14
4	2	6	2	4	0	14
3	4	1	2	4	1	12
2	0	0	0	1	6	7
1	0	0	0	0	3	3
Total	10	10	10	10	10	50

TABLE 6

Analysis of Variance for Derived Scores

Source of variation	Degrees of freedom	Sums of squares	Mean squares	F
Between fats	8	39.60	4.95	19.49
Residual	41	10.40	.254	
Total	49	50.00		$p < .01$

TABLE 7

Analysis of Covariance of Assigned and Derived Scores

Source of variation	Sums of squares (assigned scores)	Sums of cross-products	Sums of squares (derived scores)
Between fats	44.08	41.01	39.60
Residual	28.10	10.77	10.39
Total	72.18	51.78	49.99

TABLE 8

Analysis of Variance of the Derived Scores, Eliminating the Assigned Scores

Source of variation	Degrees of freedom	Sums of squares	Mean squares	F
Between fats	7	6.58	.940	6.14
Residual	41	6.26	.153	
Total	48	12.84		$p < .01$

TABLE 9

Scaling Information for All Individuals

Individuals (i)	η_i^2	Significance of discrimination between fats	Significance of variation of derived scores, eliminating unit scores
1	.7921	p < .01	p < .01
2	.3462	p < .01	p > .05
3	.3531	p < .01	p ≥ .05
4	.3023	.05 > p > .01	p > .05
5	.6129	p < .01	.05 > p > .01
6	.7241	p < .01	p < .01
7	.3675	p < .01	p > .05
8	.9327	p < .01	p < .01
9	.5031	p < .01	p > .05
10	.8147	p < .01	p < .01

TABLE 10

Mean Preferencel Scores of All Individuals for All Fats

Individuals	Fats				
	1	2	3	4	5
1	.4700	.5460	.5420	.2041	-1.7622
2	.5845	.2507	-.0919	.3502	-1.0934
3	.3678	.2466	.4890	.0485	-1.1522
4	.4828	.0496	.6780	-.8101	-.4005
5	.5975	.4179	.6315	-.1997	-1.4472
6	.5832	.5512	.5279	-.0199	-1.6422
7	.4460	.4534	.4311	-.2402	-1.0905
8	.9541	.8092	.5789	-1.0702	-1.2723
9	.5092	.5456	.6024	-.4948	-1.1623
10	.5065	.5988	.4568	.2265	-1.7884
Total	5.5016	4:4692	4.8460	-2.0055	-12.8113

TABLE 11

The Matrix MZ^{-1}

$$\begin{bmatrix}
\underline{3.8095} & .7219 & .8079 & .3331 & 1.7365 & 2.7191 & .8108 & 9.5929 & 1.1314 & 4.3287 \\
2.2703 & \underline{.5295} & .4665 & .1108 & .9861 & 1.6086 & .4563 & 5.1233 & .5943 & 2.6325 \\
2.5137 & .4616 & \underline{.5458} & .2705 & 1.1830 & 1.8123 & .5477 & 6.6805 & .7774 & 2.8376 \\
1.1178 & .1182 & .2917 & \underline{.4334} & .7639 & .9717 & .3673 & 6.7459 & .6229 & 1.2051 \\
3.2333 & .5838 & .7078 & .4238 & \underline{1.5833} & 2.3867 & .7445 & 9.8930 & 1.0841 & 3.6530 \\
3.6085 & .6788 & .7729 & .3843 & 1.7011 & \underline{2.6238} & .8011 & 10.1605 & 1.1407 & 4.1004 \\
2.4666 & .4414 & .5355 & .3330 & 1.2165 & 1.8365 & \underline{.5810} & 7.9851 & .8537 & 2.7960 \\
3.1053 & .5274 & .6950 & .6507 & 1.7200 & 2.4784 & .8496 & \underline{13.8618} & 1.3220 & 3.5246 \\
2.7042 & .4517 & .5972 & .4436 & 1.3916 & 2.0544 & .6707 & 9.7608 & \underline{1.0125} & 3.0507 \\
3.8581 & .7461 & .8128 & .3201 & 1.7486 & 2.7539 & .8191 & 9.7043 & 1.1377 & \underline{4.3967}
\end{bmatrix}$$

TABLE 12

χ^2 Test for the Dimensionality of G

Source of χ^2	Degrees of freedom	χ^2	P
First canonical variance	13	1066.3239	
Residual	27	138.1454	p < .01
Total	40	1204.4693	

as the first trial vector. Four iterations yield an essentially stable first latent root $(1/m)\ \mu = 26.0079$ and first latent vector (with arbitrary unit):

$$(.2885\quad .0520\quad .0630\quad .0426\quad .1455\quad .2179\quad .0699\quad 1.0000\quad .1038\quad .3276).$$

The χ^2 test of variance remaining after that attributable to the first root has been removed shows that the preference scores *cannot* be considered unidimensional (Table 12).

Inspection of the mean preference scores (Table 10) suggests that the departure from unidimensionality is a result of disagreement between the scores of the most reliable individual, no. 8, and those of the remaining more reliable individuals, nos. 1, 5, 6, and 10. Reliability is judged from the correlation ratios in Table 9. Individuals whose preferences show low reliability influence the test of dimensionality only slightly compared with those of high reliability. The disagreement is most marked for pastries baked from fat no. 4.

When results of this sort are encountered in practice, a careful review of the conditions of testing, the stability of the objects, and the training of the individuals is probably indicated. In the present example it would be particularly unfortunate to exclude individual no. 8, who is highly reliable, if through better controlled testing, retraining, etc., his preferences could be brought into line with those of other judges.

On the other hand, a single dimension in this example accounts for $26.0079/29.3773 = 88.53\%$ of the variance of the preference scores; hence, the error incurred by ignoring the residual variance might not be considered important in practice. In this case, individuals 1, 5, 6, 8, and 10 could be accepted as the limited panel of judges without further attention to differences among their preferences. The loss of information resulting from the exclusion of the remaining individuals is indicated by the difference in variance given by (39):

$$41(26.0079 - 23.4348) = 105.4971.$$

Taking this difference as χ^2 with 20 degrees of freedom indicates that the loss of variance is significant. It is clear, however, that the marginal increase in accuracy gained from any additional individual probably would not merit the added expense or inconvenience.

Adjusting the scale of the y_i for the selected individuals according to (40) yields

$$w_{jk} = .7720g_{1jk} + .3893g_{5jk} + .5831g_{6jk} + 2.6758g_{8jk} + .8766g_{10jk}$$

as the optimal linear estimator of the ω_{jk} with unit variance within objects and between occasions.

It is of interest to compare the coefficients of this estimator with those derived from the correlation ratios for these individuals by prior assumption of unidimensional preference scores. The resulting coefficients, brought to the same scale, are

$$.7922, \quad .3743, \quad .5708, \quad 2.6558, \quad .9015.$$

The close agreement of corresponding coefficients reflects the large proportion of variance in the preference scores accounted for by the first canonical variance.

The values for the pastries baked from the five fats are determined by applying the optimal linear estimator to the mean preference scores:

	Fat				
	1	2	3	4	5
Value:	3.933	3.596	2.922	−2.597	−7.854

These values must sum to zero.

The use of the facilities of the Agricultural Experimental Station of the University of Puerto Rico for preparing the numerical example is gratefully acknowledged.

REFERENCES

1. Bartlett, M. S. Further aspects of the theory of multiple regression. *Proc. Camb. phil. Soc.*, 1938, **34**, 33-40.
2. Bartlett, M. S. Multivariate analysis. *J. roy. statist. Soc., Suppl.*, 1947, 9, 176-190.
3. Bartlett, M. S. Internal and external factor analysis. *Brit. J. Psychol., Statist. Sect.*, 1948, 1, 73-81.
4. Bartlett, M. S. The goodness of fit of a single hypothetical discriminant function in the case of several groups. *Ann. Eugen.*, 1951, **16**, 199-214.
5. Fisher, R. A. The precision of discriminant functions, *Ann. Eugen.*, 1940, **10**, 422-429.
6. Fisher, R. A. Statistical methods for research workers. New York: Hafner, 1948.
7. Guttman, L. The quantification of a class of attributes: a theory and method of scale construction. In P. Horst et al., The prediction of personal adjustment. New York: Social Science Research Council, 1941, pp. 319-348.
8. Guttman, L. An approach for quantifying paired comparisons and rank order. *Ann. math. Statist.*, 1946, **17**, 144-163.
9. Guttman, L. Components of scale analysis. In Stouffer et al., Measurement and prediction. Princeton: Princeton Univ. Press, 1950, pp. 312-361.
10. Hartree, D. R. Numerical analysis. Oxford: Clarendon Press, 1952.
11. Holzinger, K. J. and Harman, H. H. Factor analysis. Chicago: Univ. Chicago Press, 1941.
12. Johnson, P. O. The quantification of data in discriminant analysis. *J. A. S. A.*, 1950, **45**, 65-76.
13. Kelley, T. L. Interpretation of educational measurements. Yonkers-on-Hudson: World Book, 1927.
14. Kendall, M. J. The advanced theory of statistics. London: Griffin, 1948.

15. Marriott, F. H. C. Tests of significance in canonical analysis. *Biometrika*, 1952, **39**, 58-64.

16. Maung, Khint. Discriminant analysis of Tocher's eye-color data for Scottish school children. *Ann. Eugen.*, 1941, **11**, 64-76.

17. Overman, A. and Li, J. C. R. Dependability of food judges as indicated by analysis of scores of food testing panel. *Food Res.*, 1948, **13**, 441-449.

18. Rao, C. R. Advanced statistical methods in biometric research. New York: Wiley, 1952.

19. Williams, E. J. Use of scores for the analysis of association in contingency tables. *Biometrika*, 1952, **39**, 274-289.

Manuscript received 5/28/55

Revised manuscript received 10/24/55

PSYCHOMETRIKA—VOL. 21, NO. 4
DECEMBER, 1956

AN EMPIRICAL EVALUATION OF MULTIDIMENSIONAL SUCCESSIVE INTERVALS*

SAMUEL J. MESSICK†

UNIVERSITY OF ILLINOIS

The multidimensional method of successive intervals and the method of complete triads are applied to similarity judgments of Munsell colors varying in brightness, saturation, and hue. Both methods yield configurations that correlate highly with the Munsell color structure. This validation of these scaling methods in an area of known dimensionality indicates their applicability for exploration in areas of unknown dimensionality.

Several recent multidimensional scaling procedures (**2**, **7**, **14**, **18**), based upon a Euclidean mathematical model (**19**), reveal the number and the nature of relevant dimensions in unknown areas. Empirical evaluations of the methods in areas of known dimensionality seem appropriate before applying them in in areas of unknown dimensionality. The field of color perception was selected for such validation, since the dimensions of brightness, saturation, and hue are well defined.

Richardson (**14**) applied the method of triadic combinations to judgments of similarity among some Munsell colors, which differed in saturation and brightness but were of a constant hue; the results are reported as being in essential agreement with the Munsell scheme. Torgerson (**17**) applied the complete method of triads to judgments of similarity among nine Munsell colors, which were presumably of the same red hue but differed in brightness and saturation. His two-dimensional configuration was very similar to the Munsell system. Torgerson also compared complete triads with unidimensional paired comparisons using nine gray stimuli that differed only in brightness. The methods yielded unidimensional scales that were linearly related. These results can be looked upon as a "validation" of this multidimensional scaling method. Indications are that saturation and brightness, at least, can be represented adequately by a Euclidean model (**18**).

Since the method of complete triads requires so many judgments, the task becomes prohibitive with more than ten stimuli. In order to overcome this difficulty, a multidimensional method of successive intervals (**1**, **8**) was

*This study was supported in part by the Office of Naval Research contract N6onr-270-20 with Princeton University and also in part by funds from the National Science Foundation. The research was carried out when the author was an Educational Testing Service Psychometric Fellow at Princeton University.

†Now at The Menninger Foundation.

recently developed; this method requires fewer judgments per subject and thus permits the use of a larger number of stimuli. This is an important consideration in investigations of unknown areas, for although the dimensionality does not necessarily increase with the number of stimuli, it is certainly limited by that number.

This investigation is an evaluation of the multidimensional method of successive intervals in the area of color perception. With the new procedure it is feasible to include wide stimulus variations, so the study is also a step toward a systematic multidimensional mapping of psychological color space.

If the stimuli are chosen from the Munsell color system (**11**), a comparison between the Munsell scale values and those obtained from the multidimensional procedure will permit an evaluation of the method. Since Torgerson has already applied a multidimensional scaling procedure (complete triads) to color stimuli with acceptable results (**17**), a comparison of multidimensional successive intervals with complete triads can also be looked upon as a validation procedure. One aspect of this comparison will be an investigation of the difference in difficulty of the judgments required. In complete triads, which presumably involves a much simpler judgment, the subject must decide which two of three stimuli are most similar. Successive intervals requires the subject to decide whether the members of pair *A* are more or less alike than the members of pair *B* and then to order the pairs accordingly on a continuum. The practical importance of this difference may be evaluated not only in terms of the similarity of the final structures but also in terms of the time required to make the judgments, the ease with which they are made, and their acceptability to the subject as reasonable tasks.

The Methods

The multidimensional scaling methods are techniques for estimating the interpoint distances among a set of stimuli. [For detailed discussion of these methods, see (**9**).] By requiring judgments about pairs of stimuli instead of single stimuli, these methods yield scale values which can be taken to represent the distances between the members of the pairs (**20**). The general procedure is to obtain similarity judgments among stimuli and then to scale these judgments by traditional scaling methods (**5**, **15**, **16**). The scale values obtained represent distances between stimuli, which can be analyzed to obtain the dimensionality of the space and the projections of the stimuli on a set of axes placed in the space (**10**).

In complete triads each combination of three stimuli (triad) is presented three times; subjects are asked on separate occasions whether *A* is more like *B* or *C*, whether *B* is more like *A* or *C*, and whether *C* is more like *A* or *B*. Since each of $n(n - 1)(n - 2)/6$ possible triads is presented three times, $n(n - 1)(n - 2)/2$ judgments are required from each subject. In the multidimensional method of successive intervals (**1**, **8**), on the other hand, the subject is asked only to arrange the $n(n-1)/2$ possible pairs of n stimuli into $(k+1)$ categories on a distance continuum according to the degree of similarity of the members of each pair. This procedure is a direct extension of unidimensional successive intervals (**4**, **6**), with pairs of stimuli substituted for single stimuli.

Experimental Procedures

The Stimuli and Their Mode of Presentation

Eight of the nine stimuli used by Torgerson were scaled by the method of complete triads. According to the Munsell system (**11**), these stimuli are of the same red hue but differ in brightness and saturation. Eight additional Munsell colors of the same characteristics were scaled by multidimensional successive intervals. Although these colors are of the same hue according to the Munsell system, the report of the Optical Society of America subcommittee on the spacing of the Munsell colors (**13**) indicates slight hue variations. Since the *OSA* renotations represent an attempt to obtain a closer approximation to equal-appearing intervals than the spacing in the Munsell scheme (**12**), they would seem to be superior as criteria.

The *OSA* revised designations of the 16 stimuli scaled by multidimensional successive intervals are plotted in Figure 1. The circled values are the eight stimuli also scaled by the method of complete triads. The Munsell notation designates the three dimensions of color perception as hue, value (brightness), and chroma (saturation). Theoretically steps along the value and chroma scales in the Munsell system represent equal sense distances along the respective psychological dimensions, two chroma steps representing a sense distance approximately equal to one value step.

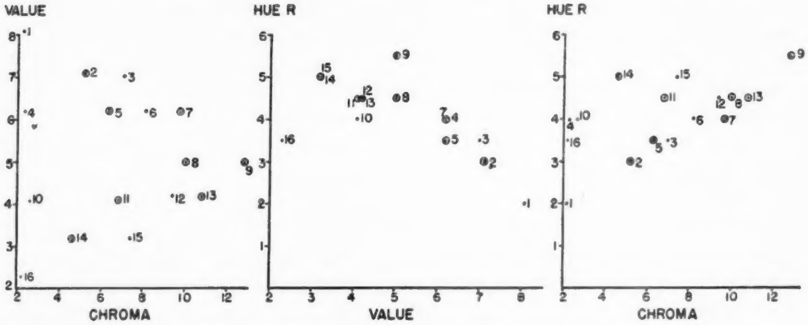

FIGURE 1
Stimulus Configuration According to OSA Revisions

For the method of complete triads, sheets of colored paper obtained from the Munsell Color Company were cut into equilateral triangles 1.5 inches on a side. Each of the 56 combinations of three different colored triangles was then mounted on a white cardboard triangle 6.25 inches on a side. For the multidimensional method of successive intervals, the colored sheets were cut into one-inch squares, and each pair of different colored squares (120 pairs for 16 stimuli) was mounted on a white 3″ × 5″ card. Each stimulus appeared as often on the left as on the right. The cards for both procedures were randomly arranged for presentation to the subjects. The stimuli were viewed against the white background of the mounts, which in turn were presented against a gray background. The lighting source was a ceiling fixture containing two GE 40-watt fluorescent daylight bulbs.

The Subjects

Forty-two subjects, 38 males and four females, took part. Color-blind persons were excluded by tests. The subjects were randomly divided into two equal groups, members of one group judging complete triads first and members of the other group judging successive intervals first.

Instructions

For the complete triads task, the subjects were to decide whether the top color on a triangular card looked more like the left one or the right one. The response of "right" or "left" was recorded by the experimenter. The subjects were also told that the colors on some triangles were very much alike, but that no triangle had colors exactly the same. They were encouraged to respond with a first impression and not to spend too much time on any one triad.

For the multidimensional successive intervals task, the subjects were to sort the stack of rectangular cards into eight piles. First they were to divide it into two approximately equal piles according to whether the colors on each card were very similar or very different. Then the subjects were to divide each of those piles into two again. If at any time a subject wished to change a card from one pile to another, he was permitted to do so. Then the subjects were asked to divide each of the four piles into two in the same way, giving eight piles. The left-hand piles were to contain cards on which the colors were the most similar, and the right-hand pile those on which the colors were the most different. In going from left to right the cards should get increasingly different. In the final step, the subjects looked through each pile to see if the cards in that pile seemed to "go together" and to make any necessary changes.

The Number of Judgments Required

The number of judgments for multidimensional successive intervals is minimal if each subject looks at each stimulus card only once and then decides in which pile it belongs. The minimum number of judgments, then, is equal to the number of stimulus cards $[n(n - 1)/2]$. Use of this procedure would have required 120 judgments from each subject in order to scale 16 stimuli. It was desired, however, to have the subject re-sort the cards after he had become familiar with them; using an expanded multidimensional successive intervals procedure, such check sorts would be obtained without the number of judgments approaching appreciably that required by complete triads. Since each of the $n(n - 1)/2$ cards was looked at four times, $2n(n - 1)$ or 480 judgments were made by each subject in order to scale 16 stimuli. If these 16 stimuli had been scaled by complete triads, however, $n(n - 1)(n - 2)/2$ or 1680 judgments would have been required.

In the complete triads section of the present experiment, each of the 56 triads was presented three times, making a total of 168 judgments required from each subject to scale eight stimuli.

Analysis and Results

The Method of Complete Triads

The raw data for the complete triads section of the study consisted of a 42 (subjects) × 168 (triads) table of responses. These data were analyzed by the complete triads procedure (**18**) to obtain a matrix of relative interstimulus distances. The general multidimensional scaling procedure (**10**) was then applied to obtain a matrix of projections, which is analogous to a factor matrix. Its rank, r, is equal to the dimensionality of the Euclidean space defined by the experimentally obtained distances, and its elements represent the projections of the stimuli on a set of r orthogonal axes placed at the centroid of the stimuli.

The matrix of projections was of rank 2. A third dimension might have been expected to represent the differences in hue. Its absence may be accounted for by the fact that the hue variations were slight compared with the range of variations in brightness and saturation. It is also possible that eight variables were too few to distinguish a third factor from error. A detailed description of the analysis and a tabulation of the data appear in (**8**). The two-dimensional structure was rotated orthogonally to approximate the Munsell dimensions. The complete triad structure and the revised Munsell structure agree closely, as can be seen in a comparison of each triads factor with the corresponding Munsell dimension (Figure 2). The present structure also had about the same degree of agreement with Torgerson's results (**17**).

The Multidimensional Method of Successive Intervals

The raw data for the successive intervals method were summarized in a 120 (stimulus-pairs) × 8 (categories) table of the number of times the ith pair of stimuli was placed in the gth category. Some of these stimulus-pairs

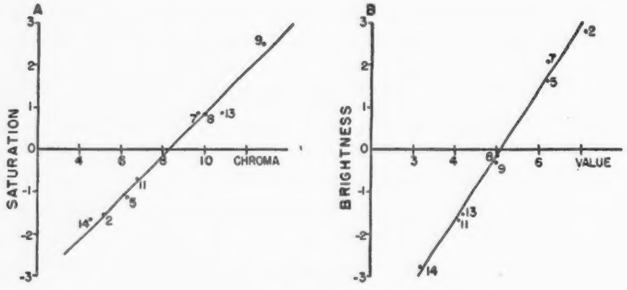

FIGURE 2
Separate Comparisons of Triad Dimensions
vs. Revised Munsell Dimensions

had been unanimously or nearly unanimously placed in one of the extreme categories, making it impossible to obtain a scale value for every pair by ordinary successive intervals procedures. If four stimuli (stimuli 6, 12, 15, and 16) were excluded from the set, however, the frequency distributions of the remaining pairs had satisfactory ranges. Thus, the set of distances among the remaining 12 stimuli was scaled by the method of successive intervals.

By the selection of two other sets of stimuli with satisfactory distributions (a set composed of the ten stimuli 1, 2, 4, 5, 6, 8, 9, 10, 12, and 15 and a set composed of the six stimuli 4, 8, 10, 12, 15, and 16), it was possible to obtain distance estimates involving all 16 stimuli. All possible distances among the 16 stimuli could not be obtained, of course, but it was hoped that there would be sufficient overlap among the three sets to allow the four excluded points to be fitted into the space defined by the set of 12 stimuli. Since some of the distances were involved in all three sets, the similarity of overlapping estimates could be evaluated. If the common distances were similarly estimated in all three groups, it would seem reasonable to attempt to fit the excluded four points into the space defined by the set of 12 stimuli.

When the three sets of stimuli were analyzed separately, the common distances were found to be estimated similarly in all three sets. Accordingly, the four excluded stimuli were fitted into the space of the 12 stimuli by a least squares criterion (**8**). The results for the set of 12 stimuli only will be considered in this paper, but it is suffice to say that this procedure located the four excluded points in positions reasonably appropriate to the Munsell scheme.

The raw data for the set of 12 stimuli, then, consisted of a 66 (stimulus-pairs) × 8 (categories) table of the number of times the ith stimulus-pair was placed in the gth category. Successive intervals procedures (**3**) yielded the inter-stimulus distances, and the general multidimensional scaling procedure (**10**) was applied to obtain a matrix of projections, which consisted of three factors. The three dimensions were rotated orthogonally to approximate the Munsell configuration. The Munsell and the successive intervals configurations are correlated to an extremely high degree. The extent of the agreement is seen in a comparison of each successive intervals factor with the corresponding Munsell dimension (Figure 3). Factor *A* corresponds to saturation, Factor *B* to brightness, and Factor *C* to hue.

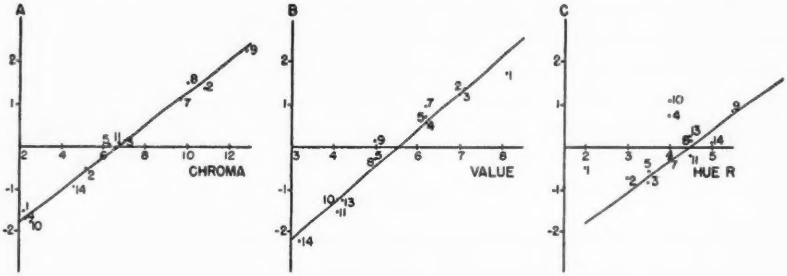

FIGURE 3
Separate Comparisons of Successive Intervals Dimensions vs. Revised Munsell Dimensions

These comparisons reveal only three widely deviant values, the hue positions obtained by successive intervals for stimuli 1, 4, and 10, which were the most unsaturated colors used. When such stimuli are compared in close proximity to stimuli with obvious hue, contrast effects might exaggerate relative inter-stimulus distances. It is also possible, however, that these deviant values indicate a slight deviation from orthogonality in the placement of the reference axes. With oblique rotation procedures these three points may be re-aligned to produce a linear plot in Figure 3 for hue vs. Factor *C*. Such an oblique orientation for Factor *C* might be considered an indication that the orthogonal axis placement in the Munsell system is somewhat in error. This is only a very tentative suggestion, but it illustrates that multidimensional scaling procedures, being based upon cross-dimensional judgments, can lead to such statements about relationships between dimensions. On the other hand, these three deviant points in Figure 3 might also be taken to indicate a slight inadequacy on the part of the Euclidean model to describe psychological color data.

A Comparison Between Complete Triads and Multidimensional Successive Intervals

A comparison between the final rotated structures obtained by complete triads and multidimensional successive intervals indicated excellent agreement. The fact that three dimensions were obtained from the successive intervals procedure and only two from complete triads is probably more a function of the stimuli than of any difference between the two techniques. Only half the number of stimuli in the successive intervals experiment were scaled by complete triads, and the variations in hue involved were slight compared with the saturation and brightness variations.

The average length of time for the 168 judgments by complete triads was 35 minutes, whereas the average time for the 480 judgments in the successive intervals task was approximately 40 minutes. It is evident that the successive intervals judgment, although presumably more difficult than the triads judgment, was made with considerable ease. Every subject accepted the task in a casual manner, without complaining about its difficulty.

Discussion

Since multidimensional scaling procedures yielded structures which correlated highly with the revised Munsell system, it would now seem reasonable to apply these procedures for purposes of exploration and discovery in areas of unknown dimensionality as well as for confirmation and modification in other areas of known dimensionality. The results of this experiment also indicate the desirability of a systematic multidimensional scaling of psychological color space and, perhaps, a modification of existing color scales. The Munsell color system and the *OSA* revised scales are based upon psychophysical investigations of each dimension separately (**12**), a procedure which does not permit the comparison of color samples across dimensions. Since the multidimensional scaling approach does allow multidimensional comparisons, a systematic application of these techniques in the color domain should permit the construction of a color solid which is based upon the relationships between as well as within dimensions.

REFERENCES

1. Abelson, R. P. A technique and a model for multidimensional attitude scaling. *Amer. Psychologist*, 1954, **9**, 319.
2. Attneave, F. Dimensions of similarity. *Amer. J. Psychol.*, 1950, **63**, 516-556.
3. Diederich, G. W., Messick, S. J., and Tucker, L. R. A general least squares solution for successive intervals. Princeton: Educational Testing Service Research Bulletin, 1955.
4. Edwards, A. L. The scaling of stimuli by the method of successive intervals. *J. Appl. Psychol.*, 1952, **36**, 118-122.
5. Guilford, J. P. The method of paired comparisons as a psychometric method. *Psychol. Rev.*, 1928, **35**, 494-506.
6. Gulliksen, H. A least squares solution for successive intervals assuming unequal standard deviations. *Psychometrika*, 1954, **19**, 117-139.
7. Klingberg, F. L. Studies in measurement of the relations between sovereign states. *Psychometrika*, 1941, **6**, 335-352.
8. Messick, S. J. The perception of attitude relationships: a multidimensional scaling approach to the structuring of social attitudes. Ph.D. thesis, Princeton Univer., 1954. Also Educational Testing Service Research Bulletin, 1954.

9. Messick, S. J. Some recent theoretical developments in multidimensional scaling. *Educ. Psychol. Measmt.*, 1956, **16**, 82-100.
10. Messick, S. J. and Abelson, R. P. The additive constant problem in multidimensional scaling. *Psychometrika*, 1956, **21**, 1-16.
11. Munsell Book of Color, Abridged Edition. Baltimore: Munsell Color Co., Inc., 1945.
12. Newhall, S. M. The ratio method in the review of the Munsell colors. *Amer. J. Psychol.*, 1939, **52**, 394.
13. Newhall, S. M., Nickerson, D., and Judd, D. B. Final report of the O.S.A. subcommittee on the spacing of the Munsell colors. *J. opt. Soc. Amer.*, 1943, **33**, 385-418.
14. Richardson, M. W. Multidimensional psychophysics. *Psychol. Bull.*, 1938, **35**, 659-660.
15. Saffir, M. A comparative study of scales constructed by three psychophysical methods. *Psychometrika*, 1937, **2**, 179-198.
16. Thurstone, L. L. Rank-order as a psychophysical method. *J. exp. Psychol.*, 1931, **14**, 187-201.
17. Torgerson, W. S. A theoretical and empirical investigation of multidimensional scaling. Ph.D. thesis, Princeton Univer., 1951.
18. Torgerson, W. S. Multidimensional scaling: I. Theory and method. *Psychometrika*, 1952, **17**, 401-419.
19. Young, G. and Householder, A. S. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 1938, **3**, 19-22.
20. Young, G. and Householder, A. S. A note on multidimensional psychophysical analysis. *Psychometrika*, 1941, **6**, 331-333.

Manuscript received 9/1/55

Revised manuscript received 11/14/55

PSYCHOMETRIKA—VOL. 21, NO. 4
DECEMBER, 1956

A NEW SCALING TECHNIQUE FOR ABSOLUTE JUDGMENTS

BURTON S. ROSNER

WEST HAVEN V. A. HOSPITAL AND YALE UNIVERSITY

The results of an experiment using the method of absolute judgments can be viewed as a matrix of conditional probabilities in which the rows represent stimuli and the columns responses. The cosine of the angle between two row vectors is a measure of the similarity of the corresponding stimuli. This cosine provides the basis for a method of scaling the stimuli. Unlike the method of paired comparisons, this new technique does not require arbitrary fixing of a unit of measurement. A numerical example is given.

In the method of absolute judgments (**2**), the experimenter selects a sample of N stimuli S_i $(i = 1, \cdots, N)$. He presents each stimulus singly to the subject; the order of presentation is randomized. The subject has available a sample of M responses R_k $(k = 1, \cdots, M)$. Each time a stimulus appears, the subject must make one and only one response. During the experiment, every stimulus is presented a number of times. Thus, the experimental data can be displayed as a matrix of conditional probabilities $p_i(k)$. This matrix has N rows, each row corresponding to a stimulus, and M columns, each column representing a response. The entry in any cell (ik) gives the conditional probability of R_k given S_i. Since the subject must make one and only one response on each trial, the sum of the probabilities in any row is 1.00.

Each row of the matrix of conditional probabilities is a vector in M-space. This fact suggests using the cosine of the angle between two row vectors as a measure of the similarity between corresponding stimuli. The cosine of the angle ϕ_{ij} between row vectors i and j, s_{ij}, is given by

$$s_{ij} \equiv \cos \phi_{ij} = \frac{\sum_k p_i(k) p_j(k)}{\sqrt{\sum_k p_i(k)^2 \sum_k p_j(k)^2}}. \tag{1}$$

Since each $p_i(k)$ is greater than or equal to zero, s_{ij} will be positive and will lie in the interval 0.00 to 1.00, inclusive. A value of 0.00 signifies that none of the responses made to S_i is ever made to S_j and vice versa. A value of 1.00 signifies that the distribution of responses to S_i is identical with the distribution for S_j. Values between the upper and lower bounds indicate the extent to which the same responses are made to both stimuli.

The measure of similarity between stimuli, s_{ij}, is in part formally identical with the Pearson product moment correlation—the product moment correlation is also the cosine of the angle between two vectors. In the latter case, the dimensionality of the space containing those vectors equals the number of joint observations in the sample. The reliability of the product moment correlation depends upon the number of joint observations. But the reliability of s_{ij} depends upon two factors: the number of observations available for determining each $p_i(k)$ and the total number of different responses R_k made to S_i and S_j . Responses made to neither stimulus do not help determine s_{ij} . Thus, a sampling distribution for s_{ij} would be very different from that for the Pearson r. Furthermore, s_{ij} assumes positive or zero values only. Therefore, s_{ij} differs from the Pearson r in several critical properties.

Since s_{ij} increases as the stimuli become more "similar" in the sense of evoking similar distributions of responses, there should be an inverse relation between the psychological scale separation of the stimuli and s_{ij} . Drs. Bert F. Green and Robert P. Abelson have pointed out how this relation can be derived. The derivation leads to a scaling technique which does not require arbitrary fixing of a unit of measurement, as is necessary in the method of paired comparisons (**4**). This new technique naturally has many features in common with previous scaling methods for absolute judgments (**1**, **2**, **3**).

Assume that the responses to S_i are normally distributed, with a mean at the actual position of the stimulus. Thus,

$$p_i(k) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left[-\frac{1}{2}\left(\frac{\bar{R}_k - \bar{S}_i}{\sigma_i}\right)^2\right] d\bar{R}_k\,, \tag{2}$$

where $\bar{R}_k$ and $\bar{S}_i$ are the scale values of R_k and S_i , respectively. To obtain $\sum_k p_i(k)p_j(k)$, use a continuous approximation:

$$\begin{aligned}\sum_k p_i(k)p_j(k) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left[-\frac{1}{2}\left(\frac{\bar{R}_k - \bar{S}_i}{\sigma_i}\right)^2\right] \frac{1}{\sqrt{2\pi}\,\sigma_j} \\ &\qquad \cdot \exp\left[-\frac{1}{2}\left(\frac{\bar{R}_k - \bar{S}_j}{\sigma_j}\right)^2\right] d\bar{R}_k \qquad (3) \\ &= \frac{1}{2\pi\sigma_i\sigma_j} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[\left(\frac{\bar{R}_k - \bar{S}_i}{\sigma_i}\right)^2 + \left(\frac{\bar{R}_k - \bar{S}_j}{\sigma_j}\right)^2\right]\right\} d\bar{R}_k\,.\end{aligned}$$

This expression is integrated by completing the square in the exponent. For convenience, let

$$A = (1/\sigma_i^2) + (1/\sigma_j^2), \tag{4}$$

$$B = (\bar{S}_i^2/\sigma_i^2) + (\bar{S}_j^2/\sigma_j^2), \tag{5}$$

and

$$D = (\bar{S}_i/\sigma_i^2) + (\bar{S}_j/\sigma_j^2). \tag{6}$$

Substituting equations (4) through (6) in (3) and completing the square,

$$\begin{aligned}\sum_k p_i(k)p_j(k) &= \frac{1}{2\pi\sigma_i\sigma_j}\int_{-\infty}^{\infty} \exp\,\{-\tfrac{1}{2}[(\sqrt{A}\,\bar{R}_k - D/\sqrt{A})^2 \\ &\qquad\qquad + B - (D/\sqrt{A})^2\}\,d\bar{R}_k \\ &= \frac{1}{2\pi\sigma_i\sigma_j}\exp\,\{-\tfrac{1}{2}[B - (D/\sqrt{A})^2]\} \\ &\qquad\qquad \cdot\int_{-\infty}^{\infty}\exp\,\{-\tfrac{1}{2}A(\bar{R}_k - D/A)^2\}\,d\bar{R}_k \\ &= \frac{1}{\sqrt{2\pi}\,\sigma_i\sigma_j}\exp\,\{-\tfrac{1}{2}[B - (D/\sqrt{A})^2]\}\,\frac{1}{\sqrt{A}}\cdot\end{aligned}\tag{7}$$

By resubstituting from equations (4) through (6) and simplifying, (7) becomes

$$\sum_k p_i(k)p_j(k) = \frac{1}{\sqrt{2\pi}\,\sqrt{\sigma_i^2 + \sigma_j^2}}\exp\left\{-\frac{1}{2}\frac{(\bar{S}_i - \bar{S}_j)^2}{\sigma_i^2 + \sigma_j^2}\right\}. \tag{8}$$

If $S_i = S_j$, then (8) reduces to

$$\sum_k p_i(k)^2 = \frac{1}{2\sqrt{\pi}\,\sigma_i}\cdot \tag{9}$$

Thus, evaluating the discriminal dispersions directly:

$$\sigma_i = \frac{1}{2\sqrt{\pi}\,\sum_k p_i(k)^2}. \tag{10}$$

Finally, taking logarithms to the base e on both sides of (8) and simplifying the resulting equation using (1) and (9),

$$\begin{aligned}(\bar{S}_i - \bar{S}_j)^2 = (\sigma_i^2 + \sigma_j^2)[\log 2 + \log \sigma_i + \log \sigma_j \\ - \log(\sigma_i^2 + \sigma_j^2) - 2\log\cos\phi_{ij}].\end{aligned}\tag{11}$$

When all discriminal dispersions are set equal to unity, (11) becomes

$$\bar{S}_i - \bar{S}_j = 2\sqrt{-\log\cos\phi_{ij}}\ . \tag{12}$$

Reduction of (11) to a computing formula yields

$$\begin{aligned}(S_i - S_j)^2 &= \frac{1}{4\pi}\left(\frac{1}{[\sum_k p_i(k)^2]^2} + \frac{1}{[\sum_k p_j(k)^2]^2}\right) \\ &\quad\cdot\left[\log 2 - \log\left(\frac{1}{[\sum_k p_i(k)^2]^2} + \frac{1}{[\sum_k p_j(k)^2]^2}\right)\right. \\ &\quad\left. - 2\log\sum_k p_i(k)p_j(k)\right].\end{aligned}\tag{13}$$

TABLE 1

Conditional Probabilities for Line Division Experiment

		Responses														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1			.03	.33	.63		.01									
2				.03	.90	.06	.01									
3					.02	.35	.55	.09	.01							
4					.03	.24	.63	.06	.04							
5	Stimuli					.01	.15	.18	.59	.07						
6									.16	.53	.28	.03				
7									.43	.49	.07	.01				
8									.03	.19	.30	.48				
9										.11	.21	.61	.07			

TABLE 2

Inter-stimulus Distances ($\bar{S}_i - \bar{S}_j$)

		S_j								
		1	2	3	4	5	6	7	8	9
	1		-0.238	-2.238	-1.884					
	2	0.238		-1.700	-1.570					
	3	2.238	1.700		-0.161	-1.598				
	4	1.884	1.570	0.161		-1.432				
S_i	5			1.598	1.432		-1.507	-1.999		
	6					1.507		-0.444	-1.908	-2.155
	7					1.999	0.444		-1.514	-1.736
	8						1.908	1.514		-0.261
	9						2.155	1.736	0.261	

TABLE 3

Successive Stimulus Scale Differences and Scale Values for the Stimuli

	d_{12}	d_{23}	d_{34}	d_{45}	d_{56}	d_{67}	d_{78}	d_{89}
1		2.000	0.354					
2	0.238		0.130·					
3	0.538	1.700		1.437				
4	0.314	1.409	0.161					
5			0.166	1.432		0.492		
6					1.507		1.464	0.247
7					1.555	0.444		0.222
8						0.394	1.514	
9						0.419	1.475	0.261
Sum	1.090	5.109	0.811	2.869	3.062	1.749	4.453	0.730
M	0.36	1.70	0.20	1.43	1.53	0.44	1.48	0.24

S_i	1	2	3	4	5	6	7	8	9
$\bar{S}_i$	0.00	0.36	2.07	2.27	3.70	5.23	5.67	7.15	7.39

Mean scale separations between stimuli can be obtained by the usual techniques and can be summed to yield scale values for the stimuli.

To illustrate the use of this scaling method, a modification of an experiment done by W. J. McGill was performed. A single subject sat before a 3-foot square black panel. An opal glass plate, 10 inches long by 4 inches wide, was mounted in the middle of the panel. Every 10 seconds, an 8-inch long horizontal line appeared on the plate for 1 second. A vertical marker on each line extended 1/2 inch above and 1/2 inch below the line. The position of the marker varied from trial to trial. The marker could appear at the midpoint of the horizontal line or 1/2, 1, 1-1/2, or 2 inches to either side of the midpoint. The subject registered judgments of the marker's position by depressing and then returning to normal one of fifteen toggle switches. The switches were mounted in a 14 by 2 by 2-inch box resting in front of the panel. Each stimulus was presented 110 times, but the first 30 trials were not used in scaling the stimuli.

Table 1 shows the matrix of conditional probabilities obtained for this experiment. The inter-stimulus distances computed from (13) appear in Table 2. Table 3 shows estimated distances between stimuli S_i and S_{i+1} . These distances were obtained by subtracting from each non-zero entry of Table 2 the corresponding entry in the next column to the right. The sums and means of the columns of Table 3 appear at the bottoms of the columns. Finally, by setting $\bar{S}_1$ at 0.00 and cumulating distances between adjacent stimuli, the scale values shown in the lower half of Table 3 are obtained. As should be expected, these values show that the subject discriminated poorly between adjacent stimuli on the extreme left or extreme right while discriminating the middle stimuli quite well.

REFERENCES

1. Attneave, F. A method of graded dichotomies for the scaling of judgements. *Psychol. Rev.*, 1949, **56**, 334-340.
2. Garner, W. R. and Hake, H. W. The amount of information in absolute judgements. *Psychol. Rev.*, 1951, **58**, 446-459.
3. Saffir, M. A comparative study of scales constructed by three psychophysical methods. *Psychometrika*, 1937, **2**, 179-198.
4. Thurstone, L. L. Stimulus dispersions in the method of constant stimuli. *J. exper. Psychol.*, 1932, **15**, 284-297.

Manuscript received 9/1/55

Revised manuscript received 11/21/55

PSYCHOMETRIKA—VOL. 21, NO. 4
DECEMBER, 1956

DETERMINATION OF THE NUMBER OF INDEPENDENT PARAMETERS OF A SCORE MATRIX FROM THE EXAMINATION OF RANK ORDERS*

JOSEPH F. BENNETT

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Two ordinal consequences are drawn from the linear multiple-factor analysis model. First, the number $R(s, d)$ of distinct ways in which s subjects can be ranked by linear functions of d factors is limited by the recursive expression $R(s, d) = R(s - 1, d) + (s - 1)\, R(s - 1, d - 1)$. Second, every set S of $d + 2$ subjects can be separated into two subsets S^* and $S - S^*$ such that no linear function of d variables can rank all S^* over all $S - S^*$, and vice versa. When these results are applied to the hypothetical data of Thurstone's "box problem," three independent parameters are found. Relations to Thurstone's suggestion for a non-correlational factor analysis are discussed.

In the introduction to *Multiple-Factor Analysis* (**3**, p. xiii), Thurstone suggested that:

> ". . . it would probably be . . . profitable to develop non-metric methods of factor analysis. An idea for such a development would be to determine the number of independent parameters of a score matrix by analyzing successive differences in rank order on the assumption that they are monotonic functions of a limited number of parameters. A score would then be regarded as merely an index of rank order, and that is essentially what we are now doing. The raw scores are transmuted into a normal distribution of unit standard deviation, and these transmuted scores are used for the correlations. Instead of dealing with the transmuted scores in this manner, one might deal with the rank orders directly . . . The actual numerical values are of secondary importance in teasing out . . . the underlying order of a new domain. . . . In putting the results to practical use, the problem would return to a metric form, with standardization and norms which call for statistical methods of the conventional kind."

Workers in the area find the suggestion interesting, but in the present state of knowledge about the relation between ordinal and metric measurement it is impossible to tell whether such a non-metric approach would have practical advantages over present correlational methods, or indeed

*The research reported here was supported jointly by the Army, Navy, and Air Force under contract with the Massachusetts Institute of Technology. The manuscript was prepared while Dr. Bennett was a Fellow of the Center for Advanced Study in the Behavioral Sciences, Stanford, California. His untimely death on May 4, 1956, at the age of 29, is regretfully announced.

whether it is even possible in theory. This paper, to forestall any initial misunderstandings, will *not* attempt to answer these questions. Its purpose is to point out two particular ordinal consequences of the factor analysis model which became apparent during the development of a generalized form (**1**) of the "Unfolding Method" of Coombs (**2**), to show how these consequences can be applied in factoring hypothetical data, and to provoke interest in the problems which would have to be solved in applying such methods to empirical data.

General Considerations

Suppose that the factor and population matrices are given. Let the factor scores of each subject be taken as the coordinates of a *subject-vector* in a *population space*, the axes of which represent a set of orthogonal factors. (A subject-vector will refer specifically to its terminus; the term *subject-point* would do as well. For example, the subspace generated by two such vectors will be regarded not as the plane spanned by them and including the origin but as the line connecting their termini and, in general, *not* containing the origin.) It is possible to represent a test in such a space as a line through the origin of the space with appropriate direction cosines relative to the axes, such that the orthogonal projections of the subject-vectors on a test-line are the subjects' scores on that test, up to a linear transformation; the order of those projections on the test-line is the order in which that test ranks the subjects. Although a ranking and its exact inverse will not be distinguished in any important way, it is convenient to give the test-line an orientation, that is, a direction in which scores are to be called *higher*.

Figure 1 represents a two-dimensional population space containing two subject-vectors, A and B. What distinguishes the class of tests which rank A over B from those which rank B over A? Construct a hyperplane, in this case a line called $H(A, B)$, through the origin and normal to the line connecting A and B. It is evident that the two subject-vectors will have coincident projections on $H(A, B)$ and distinct projections on any other line. $H(A, B)$ divides the space into two halfspaces, labeled I and II in Figure 1, such that any test-line whose orientation is from II into I will yield the ranking $A \geq B$. Such a test-line is regarded as "into I." Any test-line into II will rank $B \geq A$. (Such hyperplanes are analogous to Coombs' midpoints; the condition that they shall pass through the origin restricts all rankings to monotonicity.) A test-line coincident with $H(A, B)$ will yield the result $A = B$. In the arguments which follow the convention is adopted that such an assignment of identical scores to two subjects is not a *ranking* but the simultaneous satisfaction of two rankings, $A \geq B$ and $B \geq A$. The distinction will become important in referring to the number of distinct rankings. In Figure 1 there are two rankings, $A \geq B$ and $B \geq A$, but three possible consequences of measurement, $A > B$, $B > A$, and $A = B$.

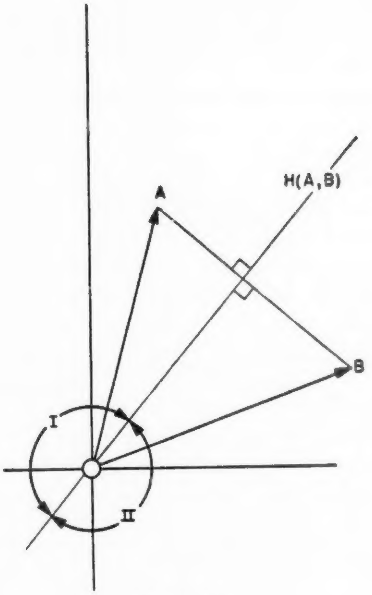

FIGURE 1

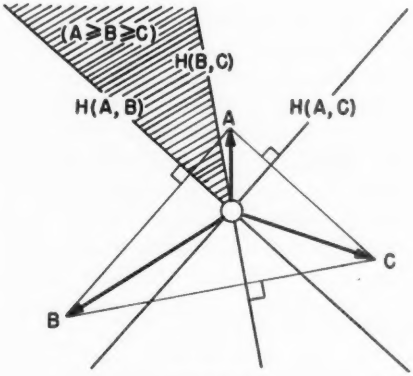

FIGURE 2

Next consider the three subject-vectors in Figure 2. Every pair of subject-vectors will generate a line through the origin which is the unique line on which their orthogonal projections coincide. The three lines so formed will together partition the space into six regions. For example, any test-line into the shaded region shown will rank *A* over *B* and *B* over *C*. These lines wholly determine the ordinal properties of the space, because the regions which they form correspond to all the possible rankings which a test could impose on the subjects.

As a final example, consider the four subject-vectors of Figure 3. There is again one line-of-coincident-projection for each pair of subject-vectors, or six such lines in all. Together they divide the space into twelve regions, each region having the property that any test into it will rank all the subjects in the same order. This means that when exactly two factors are present, there are only twelve different ways in which linear functions of those two factors can rank four subjects, and six of these rankings are exact inverses of the other six. (Note that regions which lie opposite to each other across the origin correspond to inverse rankings.) Now algebraically there are twenty-four different ways to rank four subjects; in this construction half of them are missing. This raises two questions with which to begin our investigation: why twelve rankings exactly, rather than ten or fourteen, and why these twelve?

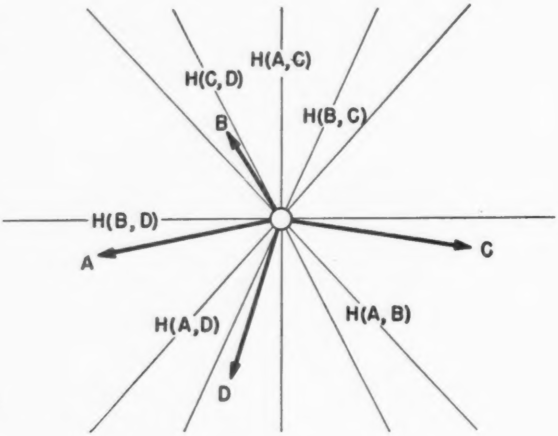

FIGURE 3

The Number of Linear Rankings

The first question is essentially: how many different ways can *s* subjects be ranked by linear functions of *d* variables? Briefly, the answer is as follows: suppose this number is already known for some given number of dimensions and subjects, and it is asked what happens if one more subject is added. One new hyperplane will be formed between each subject already present and the new subject, and each of these hyperplanes will create as many new regions as it transects, since it cuts each of them into two. The problem is, therefore, to count the number of regions which each hyperplane transects and to multiply by the number of new hyperplanes. The former number can be determined by examining the surface of each hyperplane and noting the cells into which it is divided by its intersections with the other hyperplanes; each such cell will correspond to a region through which it has passed. But how can these cells be counted? For the answer the author is indebted to Dr. Kenneth Leisenring (personal communication), who pointed out that the division of each *hyperplane* into *cells* by its intersections with other hyperplanes will be identical with the division into *regions* of the complete population space generated by one fewer subject and one fewer dimension. The reason for this identity is that each hyperplane is a legitimate subspace of the population space (that subspace from which all the variance attributable to the axis normal to the hyperplane has been extracted). All lines on its surface created by its intersections with other hyperplanes are where they ought to be if all the other subject-vectors were projected normally onto its surface and the hyperplane were treated as a population space in its own right. In this projection the two subject-vectors determining the hyperplane will be projected onto a single point, because the hyperplane was originally constructed perpendicular to the line connecting them; hence the loss of one subject as well as one dimension. So it is seen that the number $R(s, d)$ of regions or rankings created by s subjects in d-space is equal to the number present when the sth subject was added, that is, $R(s - 1, d)$, plus the number of hyperplanes added $(s - 1)$ times the number of regions which each transects, $R(s - 1, d - 1)$. The result is the recursive expression shown at the top of Table 1. This account of its derivation is rather terse because the fruit of all this labor proves to have little except theoretical interest. $R(s, d)$ quickly becomes large for moderate s and d, and so many tests are never likely to be given that the limitation on the number of different ways the subjects can be ranked becomes an important constraint on the data. It is useable directly only in the rare situation in which there are a large number of rankings of a small number of objects, as in psychophysics or in scaling experiments.

Algebraic Properties of the Permissible Rankings

The second issue is the constraint on the nature of the permissible rankings imposed by dimension, which arises in the following manner. Suppose that three subject-vectors all lie on a straight line (which does not necessarily pass through the origin) and that on this line they are in the order A, B, C. Such a configuration is illustrated in Figure 4. It is apparent that regardless of the dimensionality of the space in which these vectors occur, there are only three ways in which any linear functions of the axes can rank them: A, B, C; C, B, A; and $A = B = C$. In other words, if B is *between*

TABLE I

The Number R(s,d) of Different Ways in Which S Subjects Can Be Ranked By Linear Functions of d Factors

R(s,d) = R(s-1,d) +(s-1)R(s-1,d-1)

s	d = 2	d = 3	d = 4	d = 5
1	1	1	1	1
2	2	2	2	2
3	6 = 3!	6	6	6
4	12	24 = 4!	24	24
5	20	72	120 = 5!	120
6	30	172	480	720 = 6!
7	42	352	1,512	3,600
8	56	646	3,976	14,184
9	72	1,094	9,144	45,992
10	90	1,742	18,990	128,288
11	110	2,642	36,410	318,188
12	132	3,852	65,472	718,698
13	156	5,436	111,696	1,504,362
14	182	7,464	182,364	2,956,410
15	210	10,012	286,860	5,509,506
16	240	13,162	437,040	9,812,406
17	272	17,002	647,632	16,805,046
18	306	21,626	936,666	27,814,790
19	342	27,134	1,325,934	44,674,778
20	380	33,632	1,841,480	69,867,524
21	420	41,232	2,514,120	106,697,124
22	462	50,052	3,379,992	159,493,644
23	506	60,216	4,481,136	233,853,468
24	552	71,854	5,866,104	336,919,596
25	600	85,102	7,590,600	477,706,092
26	650	100,102	9,718,150	667,471,092
27	702	117,002	12,320,802	920,142,992
28	756	135,956	15,479,856	1,252,804,646
29	812	157,124	19,286,624	1,686,240,614
30	870	180,672	23,843,220	2,245,552,710

A and *C*, i.e., if *B* is in the interior of the interval *A–C*, no ranking based on a linear function of the coordinate axes is possible in which the subject *B* is separated from the set of subjects (*A*, *C*). Consider next the four coplanar subject-vectors in Figure 5. Note that the subject-vector *D* is in the interior of the triangle formed by vectors *A*, *B*, and *C*; consequently no linear function of the coordinate axes is possible in which the subject *D* is separated from the set of subjects (*A*, *B*, *C*). This relation may be generalized formally to an arbitrary number of dimensions in the following way: If a vector *X* is in the interior of a simplex formed by a vector-set *S*, then no linear function of the coordinate axes can generate a ranking in which the subject *X* is wholly separated from the set of subject *S*.

This relation, however true, might seem at first to have limited usefulness. What can be done with configurations like that of Figure 3, in which none of the four vectors lies in a triangle formed by the other three? The following device is proposed. Consider the subspace of dimension *d* containing a particular set of $d + 1$ subject-vectors in general position, that is, randomly scattered so that no two of them are coincident, no three form a line, no two pairs form parallel lines, etc. Consider next two such subspaces

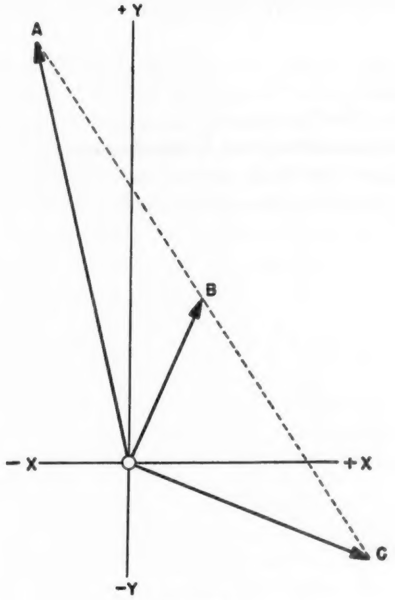

FIGURE 4

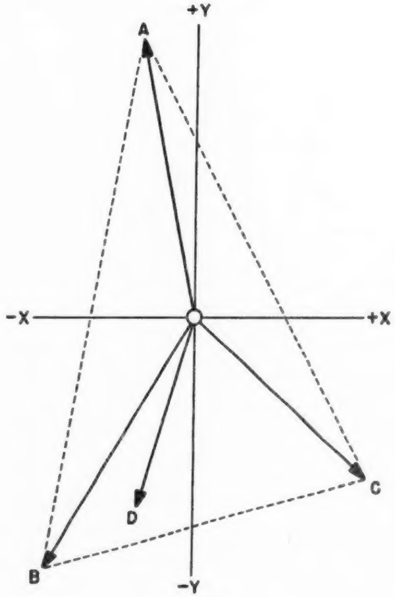

FIGURE 5

generated by two distinct subsets of subject-vectors. From familiar geometric considerations, the sum of the dimensions of two intersecting subspaces is equal to the sum of the dimensions of their union and their intersection. For example, two one-dimensional lines crossing in a plane have the whole two-dimensional plane for their union and intersect in a zero-dimensional point. In that case, one plus one equals two plus zero. Now suppose that the total number of subject-vectors in the two subsets is equal to $d + 2$, where d is the dimensionality of the population space. One of the subsets, containing n subjects, will generate an $(n - 1)$-dimensional subspace. The other, containing $d + 2 - n$ subjects, will generate a $(d + 1 - n)$-dimensional subspace. The sum of these dimensions is d. On the other hand, the dimension of their union is clearly d also, since any $d + 1$ of them are sufficient to generate the whole population space. It follows that the dimension of the intersection of the subspaces is zero; that is, they must meet in a point. Such a configuration is illustrated in Figure 6. The two subject-vectors

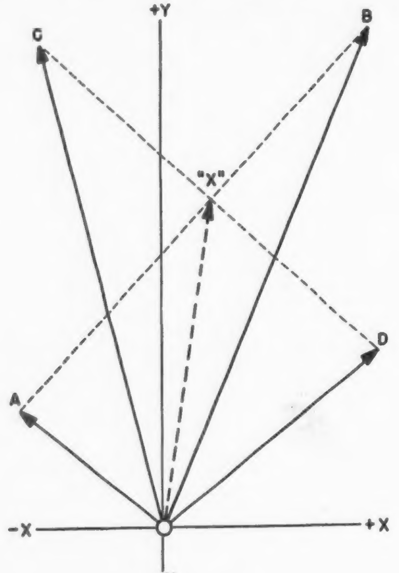

FIGURE 6

A and B generate a line $\overline{AB}$ which intersects a second line $\overline{CD}$. In this particular case, the point of intersection is between A and B on the line $\overline{AB}$ and between C and D on the line $\overline{CD}$. It is convenient to consider this point of intersection as a sort of phantom subject, X, who is a part of both subsets of subject-vectors and must obey appropriate constraints in each subset. For example, in Figure 6, no linear function of the coordinates could yield the ranking *XAB* or *ABX*. Only *AXB* and *BXA* are possible. The same is true of the line $\overline{CD}$. Furthermore, the phantom subject must continue to obey these constraints relative to each subset when the two subsets are combined in any ranking. For example, in Figure 6, the ranking *ABCD* is impossible, because it would call for the point of intersection *X* to be between *A* and *B* and at the same time to be between *C* and *D*, which is impossible. On the other hand, the order *ACBD* is possible, because it can be written *ACXBD*, placing *X* in the middle of both subsets, *AB* and *CD*. The fundamental constraint which is imposed on rankings is that no ranking is possible in which this phantom subject is called upon to be in two places at once.

There are clearly several different ways in which a given set of $d + 2$ subjects can be separated into two subsets. Of these, the most important is that which arises when the point of intersection is in the interior of the simplex in both subspaces. That is, if one of the subspaces is a line, the point of intersection is between the two points which determine the line; if one of the subspaces is a plane, the point of intersection is inside the triangle formed by the three points determining the plane; and so forth. The uniqueness of the separation of the subjects into two sets which accomplishes this leads to the following proposition: If the population space is of dimension d, then given any set S of $d + 2$ subjects, there exists a unique separation of the subjects into two subsets, S^* and $S - S^*$, such that no test can rank all the subjects in S^* over all the subjects in $S - S^*$. That is, no ranking can separate the two sets entirely, because then the phantom subject, the point of intersection, could not be inside both of them at once.

The simple cases which have been discussed, for example, that of the vector which lies in the interior of the triangle formed by three other vectors in a plane, are merely special cases of this general proposition in which one of the subsets contains but a single vector.

This result is applicable, with obvious limitations, to the problem of determining the number of independent parameters of the score matrix. The score matrix for subsets of n subjects is examined. If d independent parameters are present, at least some subsets of subjects of size $d + 1$ are separated in rank by the tests in all possible ways; and, if there are enough tests, all such subsets will be separated in all possible ways. But in every subset of $d + 2$ subjects there will be one separation which never occurs in the data, no matter how the elements of each subset are permuted within the subset. This exclusion indicates that all the rankings can be accounted for by d independent parameters.

Examples from the Box Problem

Finally, examples of the exclusion property will be drawn from the score matrix of Thurstone's box problem (**3**, p. 140), to see whether, in fact, the number of independent parameters can be determined. First, consider the scores on tests 1, 2, 4, 7, 13, 14, 18, and 19, which are known to contain only the factors X and Y. Of course when the scores are ranked the distinction among tests 1, 13, and 18 disappears at once because they are all monotonic, though non-linear, functions of X. The same is true of tests 2, 14, and 19. It is obvious that there must be more than one parameter present, because in many instances subsets of three subjects are separated in all possible ways. For example, boxes 1, 3, and 6 are variously ranked in the order 3, 6, 1 by test 4, and in the order 6, 3, 1 by test 7; no possible way of separating these three boxes into two subsets is absent. On the other hand, for every set of four boxes, there is some separation which no test imposes. For example, boxes 3 and 6 are never ranked apart from boxes 1 and 8. Box 8 is never ranked apart from boxes 1, 11, and 14. Thus two factors are sufficient to account for the score matrix. Proceeding to the analysis of all twenty tests together, it will be discovered that some sets of four subjects are separated in all possible ways. For example, boxes 1, 10, 11, and 14 are variously ranked in the order 10, 11, 14, 1 by test 17; 14, 10, 11, 1 by test 8; and 11, 14, 10, 1 by test 7. Every possible way of separating these four boxes into two subsets is represented. But in every set of five subjects a separation is discovered which does *not* occur in the data. For example, boxes 3, 5, and 7 are never separated from boxes 2 and 9. Therefore three factors are sufficient to account for all the rankings of the subjects generated by these tests.

Conclusion

The inference of factor pattern from the rankings of subjects by tests clearly has a long way to go. Both the results discussed in this paper determine dimensions in a painfully literal way. If n dimensions are needed to account for all the rankings, the technique says that there are n dimensions. But of course in any real problem there are as many dimensions as there are tests. No factor-analytic technique can hope to be useful unless some way is found to restrict the analysis to the *common* factors. In this technique, for example, it might be stipulated that *most* sets of $d + 2$ subjects shall have the exclusion property, or some such approximation. Or individual rankings might be broken down into their constituent paired comparisons, with a requirement that *nearly all* these paired comparisons shall fit the model, allowing some deviation.

Second, the determination of the number of dimensions necessary to account for the whole set of tests does not begin to settle the problem of the apportionment of the factors among the tests, that is, the questions of factor-loadings and rotation to simple structure. Thus it is still too early to say whether non-metric factor analysis of the kind proposed by Thurstone is feasible. It is the author's hope that more psychometricians will consider it worth while to find out, since the worst that can happen is that more will be learned about the underlying logic of the factor-analytic method.

REFERENCES

1. Bennett, J. F. A method for determing the dimensionality of a set of rank orders. Unpublished Ph.D. dissertation, University of Michigan, 1951.
2. Coombs, C. H. Psychological scaling without a unit of measurement. *Psychol. Rev.*, 1950, **57**, 145-158.
3. Thurstone, L. L. Multiple-factor analysis. Chicago: Univ. Chicago Press, 1947.

Manuscript received 10/24/55

Revised manuscript received 12/20/55

PSYCHOMETRIC SOCIETY

Statement of Receipts and Disbursements for Fiscal Year
Ended June 29, 1956

Receipts (Dues)

Year	Members	Student Members	
1957	1		
1956	488	58	
1955	52	14	
	541	72	
			$4075.00

Received with Dues for Corporation Publications	24.70
Proceeds of 1955 Joint Dinner with APA Division 5	.61
Overpayments	32.85
Miscellaneous	.08
Total Receipts	$4133.24

DISBURSEMENTS

Psychometric Corporation (90% of dues)	$3667.50
Psychometric Corporation (Publications)	24.70
Mimeographing and Printing	109.04
Postage	111.08
Secretarial Services	173.61
Addressing and Mailing	30.00
Bank Charges	3.48
1955 Special Program Expense	179.06
Refund of Overpayments	32.85
Telephone and Telegraph	2.31
Total Disbursements	$4333.63

BALANCE

Bank Balance, July 1, 1955	$1087.80
Receipts, 1955-56	4133.24
	$5221.04
Disbursements, 1955-56	4333.63
Bank Balance, June 29, 1956	$ 887.41

PSYCHOMETRIC CORPORATION

Statement of Receipts and Disbursements for Fiscal Year
Ended June 29, 1956

RECEIPTS

Subscriptions (less agency discounts)	$ 5362.30
Psychometric Society (90% of dues)	3667.50
Sale of Back Issues (less discounts)	1465.60
Sale of Monographs 5-8 (less discounts)	247.15
Royalties from U. Chi. Press (1954-55)	2.39
Interest on Savings Accounts	94.07
For Monographs 2, 4	5.50
Overpayments	11.00
Miscellaneous	7.65
Total Receipts	$10863.16

DISBURSEMENTS

Printing and Mailing <u>Psychometrika</u> Volume 20, No. 2, through 21, No. 1	$ 5742.15
Reprints	245.26
Stipend of Assistant Editor Volume 19, No. 3, through 20, No. 4	516.00
Secretarial Services: Editorial Office	1146.70
Secretarial Services: Business Office	303.27
Stationary and Postage	156.02
Mailing Back Issues and Monographs (12/52-3/56)	940.20
Deposited in Savings Accounts	7000.00
Refunds	38.40
Miscellaneous	42.50
Total Disbursements	$16130.50

BALANCE

Bank Balance, July 1, 1955	$ 9992.89
Receipts, 1955-56	10863.16
	$20856.05
Disbursements, 1955-56	16130.50
Bank Balance, June 29, 1956	$ 4725.55

INDEX FOR VOLUME 21

AUTHOR

In the selection of the articles to be printed in *Psychometrika*, an effort is made to obtain objectivity of choice. All manuscripts are received by one person, who first removes from each article the name of contributor and institution. The article is then sent to three or more persons who make independent judgments upon the suitability of the article submitted. This procedure seems to offer a possible mechanism for making judicious and fair selections.

Prospective authors are referred to the "Rules for Preparation of Manuscripts for Psychometrika," contained in the September, 1953 issue. Reprints of these "Rules" are available from the managing editor upon request. A manuscript which fails to comply with these requirements will be returned to the author for revision.

Authors will receive 100 reprints without covers, free of charge.

Manuscripts for publication in *Psychometrika* should be sent to

LYLE V. JONES, Managing Editor, *Psychometrika*
Dept. of Psychology, The University of Chicago
Chicago 37, Illinois

Material for review in *Psychometrika* should be sent to

JOHN B. CARROLL, Review Editor, *Psychometrika*
13 Kirkland Street
Cambridge 38, Massachusetts

The officers of the Psychometric Society for the year September 1956 to September 1957 are as follows: *President:* Hubert E. Brogden, Personnel Research and Procedures Branch, TAGO, Dept. of the Army, The Pentagon, Washington 25, D. C.; *Secretary:* Philip H. Dubois, Dept. of Psychology, Washington University, St. Louis 5, Missouri; *Treasurer:* David R. Saunders, Educational Testing Service, P. O. Box 592, Princeton, New Jersey.

The Council members, together with dates at which terms expire, are as follows: T. W. Anderson, 1959; William G. Mollenkopf, 1959; Oscar K. Buros, 1958; Alston S. Householder, 1958; Gertrude M. Cox, 1957; Harry H. Harman, 1957.

The *Psychometric Monographs Committee* is composed of Frederick B. Davis, *Chairman;* Harold Gulliksen, Paul Horst, and Frederic Kuder. Manuscripts and correspondence for this series should be addressed to

FREDERICK B. DAVIS, Chairman, Psychometric Monographs Committee,
Hunter College, 695 Park Avenue, New York 21, N. Y.

PRINTED BY THE WILLIAM BYRD PRESS, INCORPORATED, RICHMOND, VA.